Thesis for the degree of Doctor of Philosophy

Non-Gaussian Statistical Models
and Their Applications

Zhanyu Ma
马占宇

KTH Electrical Engineering

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH - Royal Institute of Technology

Stockholm 2011
行千里路
读万卷书
Abstract

Statistical modeling plays an important role in various research areas. It provides a way to connect the data with the statistics. Based on the statistical properties of the observed data, an appropriate model can be chosen that leads to a promising practical performance. The Gaussian distribution is the most popular and dominant probability distribution used in statistics, since it has an analytically tractable Probability Density Function (PDF) and analysis based on it can be derived in an explicit form. However, various data in real applications have bounded support or semi-bounded support. As the support of the Gaussian distribution is unbounded, such type of data is obviously not Gaussian distributed. Thus we can apply some non-Gaussian distributions, e.g., the beta distribution, the Dirichlet distribution, to model the distribution of this type of data. The choice of a suitable distribution is favorable for modeling efficiency. Furthermore, the practical performance based on the statistical model can also be improved by a better modeling.

An essential part in statistical modeling is to estimate the values of the parameters in the distribution or to estimate the distribution of the parameters, if we consider them as random variables. Unlike the Gaussian distribution or the corresponding Gaussian Mixture Model (GMM), a non-Gaussian distribution or a mixture of non-Gaussian distributions does not have an analytically tractable solution, in general. In this dissertation, we study several estimation methods for the non-Gaussian distributions. For the Maximum Likelihood (ML) estimation, a numerical method is utilized to search for the optimal solution in the estimation of Dirichlet Mixture Model (DMM). For the Bayesian analysis, we utilize some approximations to derive an analytically tractable solution to approximate the distribution of the parameters. The Variational Inference (VI) framework based method has been shown to be efficient for approximating the parameter distribution by several researchers. Under this framework, we adapt the conventional Factorized Approximation (FA) method to the Extended Factorized Approximation (EFA) method and use it to approximate the parameter distribution in the beta distribution. Also, the Local Variational Inference (LVI) method is applied to approximate the predictive distribution of the beta distribution. Finally, by assigning a beta distribution to each element in the matrix, we proposed a variational Bayesian Nonnegative Matrix Factorization (NMF) for bounded support data.

The performances of the proposed non-Gaussian model based methods are evaluated by several experiments. The beta distribution and the Dirichlet distribution are applied to model the Line Spectral Frequency (LSF) representation of the Linear Prediction (LP) model for statistical model based speech coding. For some image processing applications, the beta distribution is also applied. The proposed beta distribution based variational Bayesian NMF is applied for image restoration and collaborative filtering. Compared to some conventional statistical model based methods, the non-Gaussian model based methods show a promising improvement.

Keywords: Statistical model, non-Gaussian distribution, Bayesian analysis, variational inference, speech processing, image processing, nonnegative matrix factorization
List of Papers

The thesis is based on the following papers:


In addition to papers A-F, the following papers have also been produced in part by the author of the thesis:


Acknowledgements

Pursuing a Ph.D. degree is a challenging task. It takes me about four and a half years, or even longer if my primary school, middle school, high school, university, and master study are also counted. Years passed, baby is coming, good times, hard times, but never bad times.

At the moment of approaching my Ph.D. degree, I would like to thank my supervisor, Prof. Arne Leijon, for opening the door of the academic world to me. Your dedication, creativity, and hardworking nature influenced me. I benefited a lot from your support, guidance, and encouragement. Also, I would like to thank my co-supervisor, Prof. Bastiaan Kleijn, for the fruitful discussions. Special thanks also go to Assoc. Prof. Markus Flierl for the ideas that inspired me on my research.

It is a great pleasure to work with the former and current colleagues in SIP and share my research experience with you. I am indebted to Guoqiang Zhang, Minyue Li, Janusz Klejsa, Gustav Henter, and all the others for the constructive discussions about my research. I also enjoyed the teaching experience with Petko Petkov. Writing a thesis is not an easy job. I express my thanks to Dr. Timo Gerkmann, Gustav Henter, Janusz Klejsa, Nasser Mohammadiha, and Haopeng Li for proofreading the summary part of my thesis. I am grateful to Dora Söderberg for taking care of the administrative issues kindly and patiently, especially when I came to you with a lot of receipts. It would be a too long list to give every name here. Once more, I would like to thank all the SIPers sincerely.

Moving to a new country could be stressful. To Lei, David & Lili, Xi & Yi, Sha, Hailong, Qie & Bo, thank you for your generous help that makes the start of a new life so easy. To all my Chinese friends, thank you for the friendship that makes my days here wonderful and colorful.

Last but not least, I devote my special thanks to my parents and my parents-in-law for their tremendous support throughout the years. I owe my deepest gratitude to my wife Zhongwei for her support, understanding, encouragement, and love. We entered the university on the same day, received our bachelor’s and master’s degrees on the same day, flew to Stockholm on the same day, started our Ph.D. studies on the same day, got married on the same day, and will become parents on the same day. A simple “thank” is far from expressing my gratitude. I hope we can spend everyday together to make more “same days”. Love is not a word but an action. It lasts forever and ever.

Zhanyu Ma
Stockholm, December 2011
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>i</td>
</tr>
<tr>
<td>List of Papers</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>v</td>
</tr>
<tr>
<td>Contents</td>
<td>vii</td>
</tr>
<tr>
<td>Acronyms</td>
<td>xi</td>
</tr>
</tbody>
</table>

## I Summary

1. Introduction ................................. 1
2. Statistical Models ......................... 4
   2.1 Probability Distribution and Bayes’ Theorem .. 4
   2.2 Parametric, Non-parametric, and Semi-parametric Models . 6
   2.3 Gaussian Distribution .................... 7
   2.4 Non-Gaussian Distributions ............... 8
   2.5 Mixture Model ............................ 14
3. Analysis of Non-Gaussian Models ........... 16
   3.1 Maximum Likelihood Estimation ............ 16
   3.2 Bayesian Analysis ....................... 18
   3.3 Model Selection .......................... 27
   3.4 Graphical Models ....................... 28
4. Applications of Non-Gaussian Models ...... 28
   4.1 Speech Processing ....................... 28
   4.2 Image Processing ....................... 33
   4.3 Nonnegative Matrix Factorization ........ 34
5. Summary of Contributions ................... 38
   5.1 Overview ............................... 38
   5.2 Conclusions ............................ 39
References .................................... 39
II Included papers

A Bayesian estimation of beta mixture models with variational inference

1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A1
2 Beta Mixture Models and Maximum Likelihood Estimation . . A4
  2.1 The Mixture Models . . . . . . . . . . . . . . . . . . . . . . A4
  2.2 Maximum Likelihood Estimation . . . . . . . . . . . . . . . A6
3 Bayesian Estimation with Variational Inference Framework . . . A6
  3.1 Conjugate Prior of Beta Distribution . . . . . . . . . . . . . A6
  3.2 Factorized Approximation to the Parameter Distributions
    of BMM . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A7
  3.3 Extended Factorized Approximation Method . . . . . . . . A12
  3.4 Lower Bound Approximation . . . . . . . . . . . . . . . . . . A13
  3.5 Algorithm of Bayesian Estimation . . . . . . . . . . . . . . A16
  3.6 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A19
4 Experimental Results . . . . . . . . . . . . . . . . . . . . . . . . . . A20
  4.1 Synthetic Data Evaluation . . . . . . . . . . . . . . . . . . . A22
  4.2 Real Data Evaluation . . . . . . . . . . . . . . . . . . . . . . A26
5 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A31

Appendix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A31
6 Appendix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A32
  6.1 Proof of the Relative Convexity of Log-Inverse-Beta Func-
    tion (property 3) . . . . . . . . . . . . . . . . . . . . . . . . . . A32
  6.2 Relative Convexity of Pseudo Digamma Function (property 5) A34
  6.3 Approximations of the LIB Function and the Pseudo
    Digamma Function (property 4 and 6) . . . . . . . . . . . . . . A34
  6.4 Approximation of the Bivariate LIG Function (property 7) A35
Acknowledgment . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A35
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A35

B Approximating the predictive distribution of the beta distribu-
tion with the local variational method

1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . B1
2 Bayesian Estimation of the Parameters in the Beta Distribution . B3
3 Predictive Distribution of the Beta Distribution . . . . . . . . . B4
  3.1 Convexity of the Inverse Beta Function . . . . . . . . . . . . B5
  3.2 Local Variational Method . . . . . . . . . . . . . . . . . . . . B7
  3.3 Upper Bound of the Predictive Distribution . . . . . . . . . B7
  3.4 Global Minimum of the Upper Bound . . . . . . . . . . . . . B8
  3.5 Approximation of min_{u_0, v_0} F(x, u_0, v_0) . . . . . . B9
  3.6 Approximation of the Predictive Distribution . . . . . . . . B10
4 Experimental Results and Discussion . . . . . . . . . . . . . . . . . B11
5 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . B14
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A35

B Approximating the predictive distribution of the beta distribu-
tion with the local variational method

1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . B1
2 Bayesian Estimation of the Parameters in the Beta Distribution . B3
3 Predictive Distribution of the Beta Distribution . . . . . . . . . B4
  3.1 Convexity of the Inverse Beta Function . . . . . . . . . . . . B5
  3.2 Local Variational Method . . . . . . . . . . . . . . . . . . . . B7
  3.3 Upper Bound of the Predictive Distribution . . . . . . . . . B7
  3.4 Global Minimum of the Upper Bound . . . . . . . . . . . . . B8
  3.5 Approximation of min_{u_0, v_0} F(x, u_0, v_0) . . . . . . B9
  3.6 Approximation of the Predictive Distribution . . . . . . . . B10
4 Experimental Results and Discussion . . . . . . . . . . . . . . . . . B11
5 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . B14
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . A35
C Expectation propagation for estimating the parameters of the beta distribution  
1 Introduction .............................................. C1 
2 Beta Distribution and Parameter Estimation .................... C2 
  2.1 Posterior Approximation with Variational Inference ....... C3 
  2.2 Posterior Approximation With Expectation Propagation ... C4 
3 Experimental Results And Discussion .......................... C7 
  3.1 Approximation To The Posterior Distribution ............... C8 
  3.2 Model Comparison ..................................... C8 
4 Conclusion ................................................................ C10 
References .................................................................. C10 

D Vector quantization of LSF parameters with mixture of Dirichlet distributions  
1 Introduction .................................................. D1 
2 Previous Work .................................................. D4 
  2.1 LPC Parameters Representation by ∆LSF .................. D5 
  2.2 Modelling and Inter-component Bit Allocation .......... D6 
3 Decorrelation of Dirichlet Variable .............................. D7 
  3.1 Previous work on the Dirichlet variable ..................... D8 
  3.2 Three-dimensional Dirichlet Variable Decorrelation ..... D9 
  3.3 K-dimensional Dirichlet Variable Decorrelation ........ D10 
  3.4 Decorrelation of ∆LSF parameters ....................... D10 
  3.5 Computational Complexity ................................. D11 
4 Intra-component Bit Allocation and Practical Coding Scheme ... D13 
  4.1 Distortion Transformation by Sensitivity Matrix .......... D13 
  4.2 Intra-component Bit Allocation ............................. D15 
  4.3 Practical Coding Scheme .................................. D17 
5 Experimental Results and Discussion ............................ D19 
  5.1 Prerequisites ............................................... D19 
  5.2 Model Fitness ............................................. D20 
  5.3 High Rate D-R Performance ............................... D22 
  5.4 Vector Quantization Performance ......................... D23 
  5.5 Discussion .................................................. D24 
6 Conclusion ...................................................... D25 
References .................................................................. D25 

E Modelling speech line spectral frequencies with Dirichlet mixture models  
1 Introduction .................................................. E1 
2 Line Spectral Frequencies and ∆LSF ............................. E3 
  2.1 Representations ........................................... E3 
  2.2 Transformation between LSF and ∆LSF domain .......... E3 
3 Probabilistic Model Frameworks ................................. E5 
  3.1 Dirichlet Mixture Models ................................. E5 
  3.2 Parameter Estimation for DMM ......................... E6 
4 PDF-optimized Vector Quantizer .................................. E6 
  4.1 Distortion-Rate Relation with High Rate Theory ......... E7
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADF</td>
<td>Assumed Density Filtering</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion</td>
</tr>
<tr>
<td>ASRC</td>
<td>ArcSine Reflection Coefficient</td>
</tr>
<tr>
<td>BF</td>
<td>Bayes Factor</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>BMM</td>
<td>Beta Mixture Model</td>
</tr>
<tr>
<td>DAG</td>
<td>Directed Acyclic Graph</td>
</tr>
<tr>
<td>DMM</td>
<td>Dirichlet Mixture Model</td>
</tr>
<tr>
<td>DPCM</td>
<td>Differential Pulse Code Modulation</td>
</tr>
<tr>
<td>ECM</td>
<td>Expectation Conditional Maximization</td>
</tr>
<tr>
<td>EFA</td>
<td>Extended Factorized Approximation</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td>EP</td>
<td>Expectation Propagation</td>
</tr>
<tr>
<td>FA</td>
<td>Factorized Approximation</td>
</tr>
<tr>
<td>GEM</td>
<td>Generalized Expectation Maximization</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>i.i.d.</td>
<td>Independent and Identically Distributed</td>
</tr>
<tr>
<td>IS</td>
<td>Itakura-Saito</td>
</tr>
<tr>
<td>ISCA</td>
<td>International Speech Communication Association</td>
</tr>
<tr>
<td>ISF</td>
<td>Immittance Spectral Frequency</td>
</tr>
<tr>
<td>KDE</td>
<td>Kernel Density Estimator</td>
</tr>
<tr>
<td>KL</td>
<td>Kullback-Leibler</td>
</tr>
<tr>
<td>KLT</td>
<td>Karhunen-Loève Transform</td>
</tr>
<tr>
<td>LAR</td>
<td>Log Area Ratio</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>LDA</td>
<td>Latent Dirichlet Allocation</td>
</tr>
<tr>
<td>LP</td>
<td>Linear Prediction</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Prediction Coefficient</td>
</tr>
<tr>
<td>LRMA</td>
<td>Low Rank Matrix Approximation</td>
</tr>
<tr>
<td>LSF</td>
<td>Line Spectral Frequency</td>
</tr>
<tr>
<td>LVI</td>
<td>Local Variational Inference</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum A-Posteriori</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov Chain Monte Carlo</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel-Frequency Cepstral Coefficient</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>nGMM</td>
<td>non-Gaussian Mixture Model</td>
</tr>
<tr>
<td>NMF</td>
<td>Nonnegative Matrix Factorization</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>PMF</td>
<td>Probability Mass Function</td>
</tr>
<tr>
<td>RC</td>
<td>Reflection Coefficient</td>
</tr>
<tr>
<td>RGB</td>
<td>Red Green Blue</td>
</tr>
<tr>
<td>sDMM</td>
<td>super-Dirichlet Mixture Model</td>
</tr>
<tr>
<td>SI</td>
<td>Speaker Identification</td>
</tr>
<tr>
<td>SQ</td>
<td>Scalar Quantization</td>
</tr>
<tr>
<td>SV</td>
<td>Speaker Verification</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular Value Decomposition</td>
</tr>
<tr>
<td>VI</td>
<td>Variational Inference</td>
</tr>
<tr>
<td>VQ</td>
<td>Vector Quantization</td>
</tr>
</tbody>
</table>
Part I

Summary
1 Introduction

Nowadays, statistical modeling plays an important role in various research areas. Data obtained from experiment, measurement, survey, etc, can be described efficiently by a statistical model for facilitating analysis, transmission, prediction, classification, etc. The statistical modeling provides a way to connect the data with the statistics. According to some accepted theories, Cox et al. [1] defined the statistical model as “statistical methods of analysis are intended to aid the interpretation of data that are subject to appreciable haphazard variability”, McCullagh [2] simplified it as “a statistical model is a set of probability distributions on the sample space”, and Davison [3] emphasized the purpose of a statistical model as “a statistical model is a probability distribution constructed to enable inferences to be drawn or decisions made from data”. More discussions about statistical models can also be found in, for example, [4–6].

Generally speaking, a statistical model comprises one or more probability distributions. Given a set of observed data, we are free to choose any valid probability distribution to establish a statistical model for the data. Assuming the observed data are realizations of a random variable, the probability distribution is a mathematical formula that gives the probability of each value of the variable (discrete case) or gives the probability that the variable falls in a particular interval (continuous case) [7]. For the discrete variable, the Probability Mass Function (PMF) is a mathematical function to describe the probability distribution. Similarly, the Probability Density Function (PDF) is used to describe the probability distribution of the continuous variable. If the PMF or the PDF can be determined by using a parameter vector with known dimensionality, such a model is a so-called parametric model [8] (e.g., Poisson distribution, Gaussian distribution). However, if the model can be determined using a parameter vector with unknown dimensionality (i.e., the number of parameters is not set in advance and may change according to the data), it is referred as a non-parametric model [9] (e.g., the Kernel Density Estimator (KDE) [10,11], the Dirichlet process mixture model [12]). A model containing both finite dimensional and infinite dimensional parameter vectors is named as a semi-parametric model [13] (e.g., Cox proportional hazard model [14]). To summarize, the parametric model has a fixed structure, while the non-parametric model has a flexible structure. The semi-parametric model is a compromise between these two aims. In general, a statistical model describes the relation of a set of random variables to another. It could be parametric, non-parametric, or semi-parametric. Regardless of whether the statistical model is parametric, non-parametric, or semi-parametric, the model is parameterized by a parameter vector sampled from a parameter space.

A parameterized statistical model should describe the observed data efficiently, by choosing a suitable probability distribution. This choice is made depending on the properties of the data. For instance, the case of a discrete variable such as the result of tossing a coin is usually modeled by a Bernoulli distribution and a categorical distribution is used to describe a random event which takes on one of several possible outcomes [8]. An example for the case of the continuous variable, the exponential distribution describes the time difference between events in a Poisson process and the gamma distribution is frequently used as a model for waiting time. In other words, selection of a suitable probability distribution is
favorable for the efficiency of modeling the data.

Gaussian distribution (i.e., normal distribution) is the ubiquitous probability distribution used in statistics, since it has an analytically tractable PDF and analysis based on it can be derived in an explicit form [8,15–18]. Furthermore, by the technique of mixture modeling [8,19,20], the corresponding Gaussian Mixture Model (GMM) can be used to approximate arbitrary probability distributions, with a rather flexible model complexity. The research employing the Gaussian distribution and the corresponding GMM is vast (see e.g., [16,21–24]).

However, not all the data we would like to model are Gaussian distributed [25]. The Gaussian distribution has an unbounded support, while some data have a semi-bounded or bounded support. For example, the digital image pixel value is bounded in an interval, the magnitude of speech spectrum is nonnegative, and the Line Spectral Frequency (LSF) representation of the Linear Prediction Coefficient (LPC) is bounded and ordered. Recent research [26–33] demonstrated that the usage of non-Gaussian statistical models is advantageous in applications where the data is not Gaussian distributed. Common non-Gaussian distributions include, among others, beta distribution, gamma distribution, Dirichlet distribution, and Poisson distribution. The non-Gaussian distributions mentioned above all belong to the exponential family [34,35], which is a class of probability distributions chosen for mathematical convenience. Also, the distributions in the exponential family are appropriate to model the data in some applications. We restrict our attention to the non-Gaussian distribution in the exponential family in this dissertation and use the term “non-Gaussian distribution” to denote “non-Gaussian distribution in the exponential family” in the following paragraph. Similarly, the technique of mixture models can also be applied to the non-Gaussian distributions to build a non-Gaussian Mixture Model (nGMM). Even though the GMM could approximate arbitrary probability distribution when the number of mixture components is unlimited, the nGMM can model the probability distribution more efficiently than the GMM, with comparable model complexity [27–30].

An important task in statistical modeling is to fit a statistical model by estimating its parameter vector. As mentioned, a statistical model is a parameterized model. When modeling the data with a probability distribution based on some parameters, the PMF or PDF of this probability distribution can also be interpreted as a function of the parameter vector, given the observed data. In such a case, this function can be named as the “likelihood function” [36]. There are many methods to fit the parameter vector or to estimate the distribution of the parameter vector (e.g., minimum description length (MDL) estimation, Maximum Likelihood (ML) estimation). The method of finding suitable values of the parameter vector that maximizes the likelihood function is named ML estimation, which is an essential method in estimation theory. This method was introduced by Fisher in [37]. Reviews and introductions to the ML estimation can be found in, for example, [8,36,38–40]. If we treat the parameter vector as a random vector variable and assign it with a prior distribution, we can obtain the posterior distribution of the parameter vector (variable) using Bayes’ theorem [8,38,41–43]. Instead of providing a point estimate to the parameter vector, the posterior distribution provides a description of the probability distribution of the parameter vector. Such an estimate is more informative than the point estimate, since the
posterior distribution describes the possible values of the parameter vector, with a certain probability. If the point estimate is still required, the mode of the posterior distribution of the parameter vector can be traced by the so-called Maximum A-Posteriori (MAP) estimation [44,45]. In the framework of Bayesian estimation, if the prior information is non-informative, then the MAP estimate is identical to the ML estimate. Moreover, when the posterior distribution is unimodal and the amount of data goes to infinity, the Bayesian estimate converges to the ML estimate [8]. In such case, both estimates converge to the true values of the parameters.

Often, it is not practically relevant to use a single probability distribution to describe the data. Thus the mixture modeling technique is often applied in practical problems. The Expectation Maximization (EM) algorithm [45,46] and its variants, e.g., Generalized Expectation Maximization (GEM) algorithm [47] or Expectation Conditional Maximization (ECM) algorithm [48], are frequently used to carry out the ML estimation or the MAP estimation of the parameter vector. The EM based algorithms assign a point estimate to the parameter vector but do not yield a posterior distribution. To estimate the distribution of the parameter vector, the Bayesian estimation method, which involves the prior and posterior distribution as the conjugate pair, is always applied. However, the Bayesian estimation of the parameter vector is not always analytically tractable. In such case, the Variational Inference (VI) framework [8, 49–51] is frequently applied to factorize the parameter vector into sub-groups and approximate the posterior distribution. Unlike the EM based algorithms, the VI based methods result in an approximation to the posterior distribution of the parameter vector. A point estimate could also be obtained by taking some distinctive values (e.g., the mode, the mean) from the posterior distribution. With sufficiently large amount of data, the result of VI based method converges to that of the EM based methods [8]. Meanwhile, various approaches including sampling methods [52, 53] (e.g., importance sampling [8], Gibbs sampling [54]) can be applied to generate some data according to an obtained distribution.

In addition to the algebraic representation of the statistical models, the graphical model [8, 55, 56] provides a diagrammatic representation of the statistical model. With the graphical model, the relations among different variables can be easily inferred by Bayes’ theorem and the conditional independence among different variables can be obtained. By visualizing the relations, the graphical model facilitates the analysis of a statistical model.

According to the above discussion, the main features of the statistical modeling framework are, among others:

1. Usage of probability distributions, which are usually parameterized;
2. The model should be able to fit the data;
3. There exists algorithms that can be applied to estimate the parameter vector.

Utilizing an appropriate statistical model in practical applications can improve the modeling performance. For instance, in the field of pattern recognition, the statistical model based approach is the most intensively studied and applied [16]. For source coding problems, for example in speech coding, the statistical model
Summary

is applied to model the speech signal or the speech model for the purpose of data compression [57,58]. In speech enhancement, the speech, the noise, and the noisy speech are modeled by statistical models and enhancement algorithms are derived based on the estimated models [57,59]. Another application of the statistical model is found in matrix factorization. To factorize the matrix with a low rank approximation, several statistical methods, e.g., probabilistic Principal Component Analysis (PCA) [22] and Bayesian Nonnegative Matrix Factorization (NMF) [24], were proposed to give a probabilistic representation of the conventional methods.

The statistical models mentioned above are mainly based on Gaussian distribution. One can do much better if the statistical model is non-Gaussian because the true data is seldom Gaussian distributed. Thus applying non-Gaussian statistical models to the non-Gaussian data could improve the performance of the applications. This motivates us to work on non-Gaussian statistical models. The work in this dissertation focuses mainly on the following three aspects:

1. Find suitable non-Gaussian statistical models to describe data that is non-Gaussian distributed;
2. Derive efficient parameter estimation methods, including both the ML and the Bayesian estimations, for non-Gaussian statistical models;
3. Propose applications, where the non-Gaussian modeling could improve the performance.

The remaining parts is organized as follows: the concepts, the principles, and some examples of the statistical models (Gaussian and non-Gaussian) are reviewed in Section 2; the parameter estimation methods to the non-Gaussian statistical models are introduced in Section 3; the performance of the non-Gaussian statistical model is evaluated in Section 4; the contribution of this dissertation is summarized in Section 5.

2 Statistical Models

A statistical model contains a set of probability distributions to describe the data [1–3]. Based on established statistical model, inferences and prediction can be made. Usually, a statistical model describes the relation between the data and the parameters and connects them in a statistical way.

2.1 Probability Distribution and Bayes’ Theorem

A probability distribution is a formula gives the probability of a random variable with certain values [7]. For a discrete random variable $X$, a PMF $p_X(x)$ is a mathematical function to describe the probability that $X$ equals $x$, i.e.,

$$p_X(x) = \Pr(X = x),$$

(1)

where $x$ denotes any possible value that $X$ can take. The PMF satisfies the constraint that

$$p_X(x) \geq 0, \quad \sum_{x \in \Omega_X} p_X(x) = 1,$$

(2)
where $\Omega_X$ is the sample space of $X$. If we have another discrete random variable $Y$, then $p_{X,Y}(x,y)$ is the joint PMF representing the probability that $(X, Y)$ equals $(x, y)$. Furthermore, the conditional PMF $p_X(x|Y = y)$ gives the probability that $X = x$ when $Y = y$. With the sum and product rules [8], we have

$$p_X(x) = \sum_{y \in \Omega_Y} p_{X,Y}(x,y),$$

(3)

$$p_{X,Y}(x,y) = p_Y(y|X = x) p_X(X = x).$$

Also, $p_X(x)$ is named as the marginal PMF. $X$ is independent of $Y$, if

$$p_{X,Y}(x,y) = p_X(x) p_Y(y).$$

(4)

Furthermore, involving a third discrete random variable $Z$, $X$ and $Y$ are conditionally independent given $Z = z$, if

$$p_{X,Y}(x,y|Z = z) = p_X(x|Z = z) p_Y(y|Z = z).$$

(5)

Bayes’ theorem links the conditional probability $a$ given $b$ and the inverse form. Given conditional probability $p(a|b)$ and the marginal probability $p(b)$, Bayes’ theorem [8, 38, 41–43] infers the conditional probability of $b$ given $a$ as

$$p(b|a) = \frac{p(a|b)p(b)}{p(a)} = \frac{\sum p(a|b)p(b)}{\sum p(a|b)p(b)}.$$ 

(6)

With Bayes’ theorem in (6), we can obtain the PMF of $Y = y$ given $X = x$ as

$$p_Y(y|X = x) = \frac{p_{X,Y}(x,y)}{\sum_{y \in \Omega_Y} p_{X,Y}(x,y)}.$$ 

(7)

The above procedure is named as “Bayesian inference” [43, 60, 61].

When $X$ denotes continuous random variable, the PDF is used to describe the “likelihood” that $X$ takes value $x$. If the probability that $X$ falls in an interval $[x, x + \Delta]$ can be denoted as

$$\Pr(X \in [x, x + \Delta]) = \int_x^{x+\Delta} f_X(v) dv$$

(8)

and

$$\lim_{\Delta \to 0} \int_x^{x+\Delta} f_X(v) dv = f_X(x) \cdot \Delta,$$

then $f_X(x)$ is named as the PDF of $X$. As a PDF, $x$ should satisfy the following constraints

$$f_X(x) \geq 0, \ x \in \Omega_X,$$

$$\int_{x \in \Omega_X} f_X(x) dx = 1.$$ 

(10)

Similarly, given another continuous random variable $Y$, the “conditional” PDF of $X$ given $Y$ is denoted as $f_X(x|y)$ and the “joint” PDF is denoted as $f_{X,Y}(x,y)$. The sum and product rules are also valid to $X$ and $Y$ as

$$f_X(x) = \int_{y \in \Omega_Y} f_{X,Y}(x,y) dy,$$

$$f_{X,Y}(x,y) = f_Y(y|x) f_X(x).$$

(11)
Again, with Bayes’ theorem, the conditional PDF of $Y$ given $X$ is formulated as

$$f_Y(y|x) = \frac{f_X(x|y)f_Y(y)}{\int_{y \in \Omega_Y} f_X(x|y)f_Y(y)dy}. \quad (12)$$

Similarly, $X$ and $Y$ are independent if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y). \quad (13)$$

Furthermore, $X$ and $Y$ are conditionally independent given another continuous random variable $Z$ as

$$f_{X,Y}(x,y|z) = f_X(x|z)f_Y(y|z). \quad (14)$$

The PMF is aimed for discrete random variables while the PDF is aimed for continuous random variables. They have similar properties and the same inference methodology can also be applied to them. Thus the analysis, inference, and discussion in the following paragraph is based only on the continuous random variable.

### 2.2 Parametric, Non-parametric, and Semi-parametric Models

As a mathematical formula describes the connection between $X$ and $\theta$, the PDF is a function of $x$, based on the parameter $\theta$. The random variable $X$ could be either a scalar variable or a vector variable. It is the same for the parameter $\theta$. A scalar can be considered as a special case of a vector (i.e., a vector with dimensionality 1) and a vector can be considered as an extension (in the dimensionality) of a scalar. Since a PDF is usually defined for a scalar random variable (then extended for a vector random variable, if possible) and contains more than one parameter, $X$ denotes the scalar random variable and $\theta$ denotes the parameter vector in the following paragraph, except where otherwise stated. In this dissertation, a formal expression for PDF of $X$ based on $\theta$ is $f_X(x;\theta)$, where $x$ is a sample of the variable $X$. The notation $f_X(x;\theta)$ has the same mathematical expression as $f_X(x;\theta)$ but denotes the conditional PDF of the variable $X$ given the variable $\theta$, and $\theta$ is a sample of $\theta$. It would be used in Bayesian analysis.

A parametric statistical model has a fixed model complexity. In other words, the dimensionality of $\theta$ is known in advance and fixed. The parametric model is widely used in research. A scalar Gaussian distribution is a parametric model with mean and standard deviation as its parameters. A GMM contains a known (or say, pre-chosen) number of mixture components, thus it also has a fixed number of parameters. A Bayesian version of the mixture model, e.g., Bayesian GMM [62], could give zero probabilities (or slightly positive probabilities, but not distinguishable from zero) to some mixture components according to the data. However, assigning zero probability to a mixture component does not mean this mixture component “does not exist”. The component with zero probability still belongs to the whole model, but contributes nothing.
Some applications would like to have a flexible model complexity, which means the dimensionality of the parameter vector is not pre-assigned and may be changeable according to the data. This is also because a simple model is not sufficient for analysis of complicated data set. To this end, the non-parametric model [9,13,63] can be applied to describe the data efficiently and flexibly. The term “non-parametric” here means the dimensionality of the parameter vector is unknown or unbounded in advance. The number of parameters would change, which depends on the data. One of the frequently used non-parametric models is the KDE [10,11], which is applied to estimate the PDF of a variable. It is a smoothing method and could choose different kernel functions. For the purpose of Bayesian analysis, the Bayesian non-parametric methods [63,64] aim to make an informative estimation with fewer assumptions to the model. The Dirichlet process mixture model [12,65–67] is a powerful tool for modeling the data with a full flexibility on the model size. The Dirichlet process mixture model generates a discrete random probability measure for each mixture component. The activation of a new mixture component is controlled by the new data, the existing components, and a concentration parameter. This generating process is one perspective of the Chinese restaurant process [68].

A model that contains both parametric component and non-parametric component is named as a semi-parametric model [13,69]. The parametric model has a simple structure while the non-parametric model is flexible at the model complexity. The semi-parametric model combines these two features. One of the famous semi-parametric models might be the Cox proportional hazard model [14], which contains a finite dimensional parameter vector for interest and an infinite dimensional parameter vector as the baseline hazard function.

By the assistance of Bayesian framework, the model complexity in a parametric model could also be adapted. This adaptation is mainly based on assigning zero weights to some components whose contribution to the whole model can be ignored. It does not change the actual model size but limits the number of “active” parameters in the model. The model complexity is also flexible in this sense. Thus, this dissertation focuses only on how to find a suitable parametric model for the data and how to estimate the parameters efficiently. The basic idea about choosing a proper probability distribution for the data and the proposed estimation methodology can also be easily generalized to non-parametric models.

### 2.3 Gaussian Distribution

The Gaussian distribution is the most frequently used probability distribution [70,71]. There are numerous applications of the Gaussian distribution [16], the corresponding GMM [19,21] or the Gaussian process [18,72].

The Gaussian distribution has a PDF as

\[
    f_X(x; \mu, \sigma) = \mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},
\]

where \( \mu \) is the mean and \( \sigma \) denotes the standard deviation. It is easy to extend the scalar case to the multivariate case with a \( K \)-dimensional vector variable \( \mathbf{x} \).
as
\[ f(x; \mu, \Sigma) = \mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{2\pi}^K |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}, \]
(16)

with mean vector \( \mu \) and covariance matrix \( \Sigma \). The Gaussian distribution has a symmetric “bell” shape and the parameters can be easily estimated. For the Bayesian analysis, prior distributions can be assigned to the parameters and the posterior distribution of the parameters can be obtained by an analytically tractable solution [62]. This is another advantage of applying the Gaussian distribution. If the data to model are not unimodal or do not have a symmetric shape, a GMM could be applied to describe the data distribution. In principle, the GMM can model arbitrary distribution with unlimited number of mixture components. In other words, it increases the model complexity to improve the flexibility of modeling. More applications based on the Gaussian distribution can also be found in [8, 15, 17, 24, 73].

2.4 Non-Gaussian Distributions

Although the Gaussian distribution is widely applied in a lot of applications, some practical data are obviously non-Gaussian. For example, the queueing time is nonnegative, the short term spectrum of speech is semi-bounded in \([0, \infty)\), the pixel value of digitalized image is bounded in a fixed interval, and the LSF representation of the LPC is bounded in \([0, \pi]\) and strictly ordered. If we assume such type of data are Gaussian distributed, the domain of the Gaussian variable violates the boundary property of the semi-bounded or the bounded support data. Even though a general solution to such kind of problem is to apply a GMM to model the data, it requires a big amount of mixture components to describe the edge of the data’s sample space [29].

Recently, several studies showed that by applying a distribution with a suitable domain (or a mixture model of such distribution) could explicitly improve the efficiency of modeling the data and hence improve the performance (e.g., recognition, classification, quantization, and prediction) based on such modeling framework. Ji et al. [74] applied the Beta Mixture Model (BMM) in bioinformatics to solve a variety of problems related to the correlations of gene-expression. A practical Gibbs sampler based Bayesian estimation of BMM was introduced in [27]. To prevent the edge effect in the GMM, the bounded support GMM was proposed in [29] to model the underlying distribution of the LSF parameters for speech coding. As an extension of the beta distribution, the Dirichlet distribution and the corresponding Dirichlet Mixture Model (DMM) was applied to model the color image pixels in Red Green Blue (RGB) space [75]. In the area of Nonnegative Matrix Factorization (NMF), Cemgil et al. [31] proposed a Bayesian inference framework for NMF based on Poisson distribution. A exponential distribution based Bayesian NMF was introduced in [32] for recorded music. These applications focus on capturing the semi-bounded or bounded support properties of the data and apply an appropriate distribution to model the data so that the model complexity is decreased as well as the application performance is improved, compared to some Gaussian distribution based methods. Thus, it is more convenient and efficient to utilize a distribution which has a suitable domain and clear
To this end, this dissertation analyzes several non-Gaussian distributions, e.g., the beta distribution, the Dirichlet distribution, and the gamma distribution. Also, it studies the parameter estimation methods to some non-Gaussian distributions. Some applications in the area of speech processing, image processing, and NMF are evaluated to validate the proposed methods. These contents are covered in the attached papers A-F.

One of the obvious and main differences between the Gaussian and non-Gaussian distributions is the sample space of the variable \( X \). In the remaining parts of this section, some non-Gaussian distributions will be introduced. The methods of estimating the parameters and the corresponding applications will be introduced in section 3 and 4, respectively.

**Beta Distribution**

The beta distribution is a family of continuous distribution defined on the interval \([0, 1]\) with two parameters. The PDF of the beta distribution is

\[
    f_X(x; u, v) = \frac{1}{\text{Beta}(u, v)} x^{u-1} (1 - x)^{v-1}, \quad u, v > 0, \tag{17}
\]

where \( \text{Beta}(u, v) \) is the beta function defined as \( \text{Beta}(u, v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)} \) and \( \Gamma(\cdot) \) is the gamma function defined as \( \Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt \). The mean value, the mode, and the variance are

\[
    \bar{X} = E[X] = \frac{u}{u + v}, \quad \text{Mode}(X) = \frac{u - 1}{u + v - 2}, \quad \text{if} \quad u, \ v > 1, \tag{18}
\]

\[
    \text{Var}(X) = E \left[ (X - \bar{X})^2 \right] = \frac{uv}{(u + v)^2(u + v + 1)}
\]

respectively. The differential entropy of the beta variable is

\[
    H(X) = \ln \text{Beta}(u, v) - (u - 1) \psi(u) - (v - 1) \psi(v) + (u + v - 2) \psi(u + v), \tag{19}
\]
where $\psi(\cdot)$ is the digamma function defined as $\psi(z) = \frac{\partial \ln \Gamma(z)}{\partial z}$. The beta distribution has a flexible shape, it could be symmetric, asymmetric, or convex, which depends on the parameters $u$ and $v$. These two parameters play similar roles in the beta distribution. Examples of the beta distribution can be seen in Fig. 1. For the bounded support data beyond the interval $[0, 1]$, it is straightforward to shift and scale the data to fit in this interval.

The multivariate beta distribution can be obtained by cascading a set of beta variables together, i.e., each element in the $K$-dimensional vector variable $\mathbf{x}$ is a scalar beta variable. The PDF of the multivariate beta distribution is expressed as

$$f_{\mathbf{x}}(\mathbf{x}; u, v) = \prod_{k=1}^{K} \text{Beta}(x_k; u_k, v_k)$$

$$= \prod_{k=1}^{K} \frac{\Gamma(u_k + v_k)}{\Gamma(u_k)\Gamma(v_k)} x_k^{u_k-1}(1-x_k)^{v_k-1}, \quad u_k, v_k > 0. \tag{20}$$

Unlike the multivariate Gaussian distribution, the covariance matrix of the multivariate beta distribution is diagonal. Thus the correlation among the elements in the vector variable can not be described by a single multivariate beta distribution. In such case, the BMM is applied and the correlation is handled by the mixture modeling.

The beta distribution is usually applied to model events that take place in a limited interval and is widely used in financial model building, project control systems, and some business applications [76]. For the purpose of Bayesian analysis, it is often used as the conjugate prior to the Bernoulli distribution [76–78]. In the field of electrical engineering, it is not usual to model the distribution of the data directly with the beta distribution. The reason that the beta distribution (or the corresponding BMM) has not received so much attention might be due to the difficulties in parameter estimation, where a closed-form solution does not exist and some approximations are required [27, 28]. To solve this problem, the EM algorithm for the ML estimation of the parameters in the BMM was proposed in [79]. To make a more informative estimate, a practical Bayesian estimation method to the BMM through the Gibbs sampling was introduced in [27]. In contrast to the method based on sampling, we proposed an analytically tractable solution to obtain the posterior distribution of the BMM in [28], i.e., the attached paper A. A brief introduction to the closed-form solution will be found in section 3.2 and more details are given in the attached paper A.

Furthermore, we also proposed a new NMF strategy, the beta-gamma (BG-NMF), for Low Rank Matrix Approximation (LRMA) to the bounded support data [33]. Each element in the matrix is modeled by a beta distribution and then the parameter matrices are factorized. More details about the BG-NMF can be found in the attached paper F.

**Dirichlet Distribution**

The Dirichlet distribution is the conjugate prior to the multinomial distribution, which is parameterized by a vector parameter $\alpha$ and each element in the parameter vector variable is positive. The $K$-dimensional vector variable $\mathbf{x}$ in the
Dirichlet distribution is following two constraints: 1) each element in the vector variable is positive and 2) the sum of the elements is equal to one. Thus a $K$-dimensional Dirichlet variable has $K - 1$ degrees of freedom. The PDF of the Dirichlet distribution is defined as

$$f_k(x; \alpha) = \text{Dir}(x; \alpha) = \frac{\Gamma \left( \sum_{k=1}^{K} \alpha_k \right)}{\prod_{k=1}^{K} \Gamma \left( \alpha_k \right)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}. \quad (21)$$

The mean, mode, and covariance matrix of the Dirichlet distribution are

$$\bar{X}_k = \frac{\alpha_k}{\alpha_0}, \quad \text{Mode} \left( X_k \right) = \frac{\alpha_k - 1}{\alpha_0 - K}, \quad \text{if } \alpha_k > 1,$$

$$\text{Var}(X_k) = \frac{\alpha_k}{\alpha_0^2 (\alpha_0 + 1)}, \quad \text{Cov}(X_l, X_k) = \frac{-\alpha_l \alpha_k}{\alpha_0^2 (\alpha_0 + 1)} l \neq k, \quad (22)$$

respectively and $\alpha_0 = \sum_{k=1}^{K} \alpha_k$. The differential entropy of the Dirichlet variable is

$$H(X) = \sum_{k=1}^{K} \ln \left( \Gamma \left( \alpha_k \right) \right) - \ln \left( \Gamma \left( \sum_{k=1}^{K} \alpha_k \right) \right) + (\alpha_0 - K) \psi(\alpha_0) - \sum_{k=1}^{K} (\alpha_k - 1) \psi(\alpha_k). \quad (23)$$

Fig. 2 shows an example of a three-dimensional Dirichlet distribution with different view angles. It can be recognized that the elements in the Dirichlet variable are negatively correlated, which means if we increase one element then another (or some other) element(s) should be decreased correspondingly. This property plays an important role in the mixture modeling, where the Dirichlet distribution is used to model the prior/posterior distribution of the weighting factors to the mixture components. In the Bayesian analysis, if the $k$th mixture component has a very small weighting, i.e., the corresponding parameter $\alpha_k$ in the Dirichlet distribution is relative small in the whole parameter vector, this mixture component has very little effect to the whole model, and can thus be discarded.
the Dirichlet process mixture model [12, 65–67], the Dirichlet process is considered as an infinite dimensional Dirichlet distribution, where only a limited number of mixture components are activated by the observed data. Also, it is still possible to activate a new mixture component with the upcoming data.

Two more distinguished properties of the Dirichlet variable should also be mentioned here: the aggregation property [80] and the neutrality [81]. By the aggregation property of the Dirichlet variable, a $K$-dimensional Dirichlet vector variable can be aggregated to a $K - 1$ dimensional vector variable as

$$X' = [X_1, \ldots, X_k + X_{k+1}, \ldots, X_K]^T \sim \text{Dir}(x', \alpha'),$$

with $\alpha' = [\alpha_1, \ldots, \alpha_k + \alpha_{k+1}, \ldots, \alpha_K]^T$. As the elements in the Dirichlet variable are permutable, any two elements can be added together and then a new Dirichlet variable with new parameter vector is obtained. The above procedure can be done iteratively and a beta variable $[X, 1 - X]^T$ can be obtained finally. The neutrality of the Dirichlet variable was introduced by Connor et al. [81] to study proportions in biological data. The neutrality of a $K$-dimensional Dirichlet variable can be expressed as

$$f_{X_k \leq h}(x_k, \frac{x_1}{1 - x_k}, \ldots, \frac{x_K}{1 - x_k}) = f_{X_k}(x_k)f_{X \leq h}(\frac{x_1}{1 - x_k}, \ldots, \frac{x_K}{1 - x_k}),$$

which means if we take an arbitrary element from the Dirichlet variable and normalize the remaining elements, the selected element is independent of the remaining normalized ones.

The Dirichlet distribution is usually used as the conjugate prior to the multinomial distribution in Bayesian analysis [82, 83]. Instead of applying the Dirichlet distribution as the prior/posterior distribution, Bouguila et al. [75, 84] applied the DMM directly to the color image pixel value, where the pixel value in the RGB space is considered as a three-dimensional vector and normalized to satisfy the constraints of the Dirichlet variable. A generalized DMM was applied for eye modeling in [85]. Blei et al. proposed the Latent Dirichlet Allocation (LDA) model for collections of discrete data such as text corpora [26, 86]. For speech coding, we utilized the boundary and ordering properties of the LSF representation to introduce a new representation named $\Delta$LSF [30, 87]. The $\Delta$LSF satisfies the constraints of the Dirichlet variable, thus a DMM is used to model the underlying distribution of the $\Delta$LSF vectors. Furthermore, based on the obtained DMM, the aggregation property, and the neutrality, we proposed a PDF-optimized Vector Quantization (VQ) to quantize the Linear Prediction (LP) model. A brief introduction will be introduced in section 4.1 and full details are proposed in the attached papers D and E.

Although in some literature [75, 84] the Dirichlet distribution is considered as the multivariate generalization of the beta distribution, we do not follow this terminology here. A $K$-dimensional Dirichlet variable satisfies the constraint of unit summation, thus the sample space is a standard $K - 1$ simplex (see (21)). However, each element in the vector variable of multivariate beta distribution (please refer to (20)) has the support $[0, 1]$, which differs from the elements in the Dirichlet variable. For example, if we set $K = 3$, then the sample space of the Dirichlet variable in (21) is a simplex (a triangle plane) while the support of
(a) Dirichlet distribution.  
(b) Multivariate beta distribution.

Figure 3: Domain comparison of three-dimensional Dirichlet distribution and three-dimensional multivariate beta distribution.

the multivariate beta distribution in (20) is within a unit cube. A comparison of the sample spaces of the three-dimensional Dirichlet and the three-dimensional multivariate beta distribution is illustrated in Fig. 3. In this dissertation, we consider the multivariate beta distribution and the Dirichlet distribution as two different cases, thus these two types of distributions can be applied for different data. The ML estimation of the Dirichlet distribution and the corresponding DMM can be found in, e.g., [75, 88, 89]. Also, the ML estimation of the multivariate beta distribution/BMM and the Bayesian estimate of multivariate beta distribution/BMM can be found in [79] and [28], respectively. To demonstrate the effects of different sample spaces, we also proposed a multivariate BMM based VQ and compared it with the DMM based VQ in [30, 87, 89].

Gamma Distribution

The gamma distribution is well-known for modeling waiting times [90]. As the variable in the gamma distribution is nonnegative, Dat et al. [91] also applied the gamma distribution for modeling the speech power in speech enhancement. The PDF of the gamma distribution is represented as

$$f_X(x; \nu, \beta) = \text{Gam}(x; \nu, \beta) = \frac{\beta^\nu x^{\nu-1} e^{-\beta x}}{\Gamma(\nu)}, \quad \beta, \nu > 0,$$

where $x$ is a nonnegative variable, $\beta$ is the inverse scale parameter, and $\nu$ is the shape parameter. The mean value, the mode, and the variance are

$$\mu = E[X] = \frac{\nu}{\beta},$$

$$\text{Mode}(X) = \frac{\nu - 1}{\beta}, \quad \text{if } \nu > 1,$$

$$\text{Var}(X) = E \left[ (X - \mu)^2 \right] = \frac{\nu}{\beta^2},$$

$$\nu > 1,$$
respectively. If \( \nu \) is an integer, the gamma distribution is representing a sum of \( \nu \) independent exponential variables, each of which has the same rate parameter \( \beta \). Furthermore, if \( \nu = 1 \), the gamma distribution is exactly the same as an exponential distribution. Fig. 4 shows examples of the gamma distribution with different parameter pairs.

In the field of Bayesian analysis, the gamma distribution is often used as the conjugate prior to many distributions, such as the exponential distribution, the Gaussian distribution with fixed mean, and the gamma distribution with known shape parameter. In [92], the gamma distribution was used as the prior distribution for the precision parameter \( \frac{1}{\sigma^2} \). Hoffman et al. [32] proposed a NMF scheme for recorded music, in which the gamma distribution is used as the conjugate prior to the exponential distribution. In this dissertation, we mainly consider the gamma distribution as the prior distribution to the parameters in the beta distribution, because the variable in the gamma distribution is nonnegative. Even though the gamma distribution is not a conjugate prior to the beta distribution, with some approximations and by the strategy of Variational Inference (VI), we can still obtain a conjugate pair and derive an analytically tractable solution for the Bayesian analysis of beta distribution. In the attached paper A and F, we presented the details of how to apply the gamma distribution as the prior distribution to the parameters in the beta distribution.

2.5 Mixture Model

Most of the statistical distributions are unimodal, which means the distribution has a single mode. However, the data in real applications are usually multimodally distributed. To handle the multimodality of the data and to describe the data distribution flexibly, the mixture modeling technique was introduced [8, 19, 20]. The mixture model PDF is a linear combination of several single density functions (mixture components), where each mixture component is assigned with a positive weighting factor. The sum of the weighting factors is equal to 1 so that a mixture

\[ \text{Gam}(x; \nu, \beta) \]

\[ \text{Gam}(x; \nu, \beta) \]

\[ \text{Gam}(x; \nu, \beta) \]

\[ (a) \ \nu = 1, \beta = 2 \]

\[ (b) \ \nu = 2, \beta = 2 \]

\[ (c) \ \nu = 4, \beta = 2 \]

Figure 4: Examples of gamma distribution.

\(^1\)In the original paper, the inverse-gamma distribution was used as the prior distribution for the variance parameter. As the gamma and inverse-gamma distributions are an inverse pair, the variance and the precision parameters are also an inverse pair, these two statements are equivalent.
model is also a normalized PDF. The mathematical expression for a mixture density with $I$ mixture components is

$$F_X(x; \theta) = \sum_{i=1}^{I} \pi_i f_X(x; \theta_i), \quad 0 < \pi_i < 1, \quad \sum_{i=1}^{I} \pi_i = 1,$$  \hspace{1cm} (28)$$

where $\theta = [\pi, \theta_1, \ldots, \theta_I]$ denotes all the parameter vectors, $f_X(x; \theta)$ can be any parameterized distribution, and $\theta_i$ is the parameter vector for the $i$th mixture component.

The most frequently used mixture model is the Gaussian Mixture Model (GMM) [19, 21, 62, 73], which contains several Gaussian distributions. When applying the non-Gaussian statistical models, the concept of mixture modeling can be easily extended to a mixture model with several non-Gaussian distributions. The Beta Mixture Model (BMM) was used in several applications [27, 28, 74, 79]. Also, the mixture of Dirichlet distributions were applied to model image pixels [75], to the eye modeling [85], to the LSF quantization [30, 89], etc. The Bayesian gamma mixture model was applied for radar target recognition in [93]. Fig. 5 shows a GMM and a BMM.

In this section, we introduced some basic concepts about statistical models, probability distributions, and probability density functions. Instead of the classic Gaussian distribution, we focused on the non-Gaussian distributions, e.g., the beta distribution, the Dirichlet distribution, and the gamma distribution, which are efficient for modeling data with “non-bell” shape and semi-bounded/bounded support. Finally, the technique of mixture modeling was briefly summarized. In the next section, the methods of analyzing the non-Gaussian statistical models will be presented.
3 Analysis of Non-Gaussian Models

When applying non-Gaussian statistical models to describe the distribution of the data, we usually assume that one or more parameterized models fit the data. Key analysis tasks for non-Gaussian statistical models include parameter estimation, derivation of the predictive distribution, and model selection. In this section, we first introduce different methods for estimating the parameters, which cover ML estimation and Bayesian estimation. Then we present how to select a suitable statistical model according to the data. Finally, graphical models are introduced as a supplemental tool for analyzing non-Gaussian statistical models.

3.1 Maximum Likelihood Estimation

Maximum Likelihood (ML) estimation is a widely used method for estimating the parameters in statistical models. Assuming that the variable $X$ is distributed following a PDF as

$$X \sim f_X (x; \theta),$$  \hfill (29)

then given a set of i.i.d. observations $X = \{x_1, \ldots, x_N\}$, the joint density function can be expressed as

$$f (X; \theta) = \prod_{n=1}^{N} f_X (x_n; \theta).$$  \hfill (30)

If we interpret the PDF in (30) as the likelihood function of the parameter $\theta$, then the ML estimate of the parameter is denoted as

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} \prod_{n=1}^{N} f_X (x_n; \theta).$$  \hfill (31)

Usually, it is more convenient to work with the logarithm of the likelihood function, i.e., the log-likelihood function, as

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} \sum_{n=1}^{N} \ln f_X (x_n; \theta),$$  \hfill (32)

which is equivalent to maximizing the original likelihood function.

If the log-likelihood function has global maxima and it is differentiable at the extreme points, we can take the derivative of the log-likelihood function with respect to the parameter vector and then set the gradient to zero as

$$\frac{\partial \ln L (\theta; X)}{\partial \theta} = \sum_{n=1}^{N} \frac{\partial \ln f_X (x_n; \theta)}{\partial \theta} = 0.$$  \hfill (33)

Solving the above equation could then lead to ML estimates of the parameters.

For the Gaussian distribution, the ML estimates of the parameter $\theta = [\mu, \sigma]^T$ can be simply expressed in a closed-form solution as

$$\hat{\mu}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n, \quad \hat{\sigma}_{\text{ML}} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu}_{\text{ML}})^2}.$$  \hfill (34)
However, for some non-Gaussian distributions, the ML estimates do not have an analytically tractable solution. For example, substituting the beta PDF in (17) into (33) will yield an expression\(^2\)

\[
\begin{bmatrix}
\psi(u + v) - \psi(u) + \frac{1}{N} \sum_{n=1}^{N} \ln x_n \\
\psi(u + v) - \psi(v) + \frac{1}{N} \sum_{n=1}^{N} \ln (1 - x_n)
\end{bmatrix} = 0. \tag{35}
\]

As the digamma function \(\psi(\cdot)\) is defined through an integration, a closed-form solution to (35) does not exist. Numerical methods, e.g., Newton’s method\(^9\), are thus required to solve this nonlinear problem. In\(^7\), the Newton-Raphson method was applied to calculate the solution numerically. This works well in practical problems.

When computing ML estimate of the parameters in the mixture model, the Expectation Maximization (EM) algorithm\(^8\) is generally applied. Recall the expression of the mixture model in (28) and assume we have a set of i.i.d. observations \(X = \{x_1, \ldots, x_N\}\). A \(I\)-dimensional indication vector \(z_n\) is introduced for each observation \(x_n\). The indication vector \(z_n\) has only one element equals 1 and the remaining elements equal 0. If the \(i\)th element of \(z_n\) equals 1, i.e., \(z_{in} = 1\), we assume that the \(n\)th observation was generated from the \(i\)th mixture component. Apparently, the indication vector \(z_n\) follows a categorical distribution with parameter vector \(\pi = [\pi_1, \ldots, \pi_I]^T\) as

\[
f_{Z_n}(z_n; \pi) = \prod_{i=1}^{I} \pi_{izn}^i. \tag{36}
\]

In the expectation step, the expected value of \(z_{in}\) is calculated with the current estimated \(\pi\) as

\[
\mathbb{E}[Z_{in}|X, \theta] = \frac{\pi_i f_X(x_n; \theta_i)}{F_X(x_n; \theta)}. \tag{37}
\]

In the maximization step, the weight factor \(\tilde{\pi}_i\) is calculated as

\[
\tilde{\pi}_i = \frac{1}{N} \sum_{n=1}^{N} Z_{in}. \tag{38}
\]

The parameter vector estimate of \(\theta_i\) can be obtained by taking the derivative of the logarithm of (28) with respect to the parameter vector and setting the gradient equal to zero as

\[
\frac{\partial}{\partial \theta} \sum_{n=1}^{N} \ln F_X(x_n; \theta) = \sum_{n=1}^{N} \pi_i \frac{\partial f_X(x_n; \theta_i)}{\partial \theta_i} = \sum_{n=1}^{N} \pi_i f_X(x_n; \theta_i) \frac{\partial \ln f_X(x_n; \theta_i)}{\pi_{in}} \tag{39}
\]

\[= 0.\]

\(^2\)To prevent infinite numbers in the practical implementation, we assign \(\varepsilon_1\) to \(x_n\) when \(x_n = 0\) and \(1 - \varepsilon_2\) to \(x_n\) when \(x_n = 1\). Both \(\varepsilon_1\) and \(\varepsilon_2\) are slightly positive real numbers.
This equation is a weighted-sum version of (33). Thus for GMM, we can still have a closed-form solution in the maximization step. Similar to some non-Gaussian distributions, a mixture model which contains non-Gaussian distributions as mixture components typically does not have analytically tractable solutions. Thus we can apply the same strategies for ML estimation of non-Gaussian distributions to carry out the maximization step. By performing the expectation step and the maximization step iteratively, we obtain an EM algorithm for a mixture of non-Gaussian distributions. The EM algorithm for the BMM and the DMM were introduced in [79] and [75], respectively.

Note that the EM algorithm may not simplify the calculations, if the maximization step is not analytically tractable [95]. Some variants of the EM algorithm, e.g., GEM [47] or ECM algorithm [48], can be applied to overcome such kind of problem. Aside from this drawback, the EM algorithm is not guaranteed to reach a global maximum when the log-likelihood function is non-convex [95]. Then the EM algorithm may only find a local maximum.

Another drawback of ML estimation is the problem of overfitting [7]: if the amount of observations is relatively small compared to the number of parameters, the model estimated may have a very poor predictive performance. We may utilize cross-validation [96, 97], regularization [98, 99], or Bayesian estimation [6, 8, 100] techniques to avoid such problems. In the next section, we focus on Bayesian estimation of the parameters in non-Gaussian statistical models.

3.2 Bayesian Analysis

In ML estimation, the parameter vector $\Theta$ is assumed with fixed but unknown value. ML estimation thus only gives a point estimate of the parameters. If we consider the parameter vector $\Theta$ a vector-valued random variable with some distribution, then a Bayesian estimate [6, 8, 100] of the parameter vector can be made. Given a set of observations $X = \{x_1, \ldots, x_N\}$ and assuming that the parameter vector is a random variable, the likelihood function of $\Theta$ is

$$L(\Theta) = f(X|\Theta) = \prod_{n=1}^{N} f_X(x_n|\Theta).$$

(40)

Meanwhile, the prior distribution of $\Theta$ is assumed to have a parameterized PDF following

$$\Theta \sim f_{\Theta}(\Theta; \omega_0),$$

(41)

where $\omega_0$ is called the hyperparameter of the prior distribution. By combining Bayes’ theorem in (6) with (40) and (41), we can derive the posterior density function of the variable $\Theta$ given the observations $X$, as

$$f_{\Theta}(\Theta|X; \omega_N) = \frac{f(X|\Theta) f_{\Theta}(\Theta; \omega_0)}{\int_{\Theta \in \Omega_{\Theta}} f(X|\Theta') f_{\Theta}(\Theta'; \omega_0) d\Theta'} \propto f(X|\Theta) f_{\Theta}(\Theta; \omega_0),$$

(42)

where $\omega_N$ is the hyperparameter vector of the posterior distribution but may refer to a sample space different from that of $\omega_0$. By (42), the posterior distribution of the variable $\Theta$ given the observation $X$ can be obtained. The posterior describes the shape and the distribution of the parameter. Thus Bayesian estimation is
3. ANALYSIS OF NON-GAUSSIAN MODELS

more informative than ML estimation. A point estimate can also be made based on the posterior distribution. For example, the posterior mode is

$$\hat{\theta}_{MAP} = \arg \max_\theta f(\theta|X;\omega_N),$$

(43)

which is the Maximum A-Posteriori (MAP) estimate of the parameter. Also, we can estimate the posterior mean, the posterior variance, etc.

Conjugate Priors and the Exponential Family

If the prior distribution in (41) and the posterior distribution in (42) have the same mathematical form, then the prior distribution is called the conjugate prior to the likelihood function and the prior distribution and the posterior distribution are said to be a conjugate pair [8].

Among all distributions, the distributions that belong to the exponential family [8, 34, 35, 56] have conjugate priors. The exponential family is defined by [8]

$$f_X(x|\theta) = h(x) q(\theta) e^{\theta^T u(x)},$$

(44)

where $h(\cdot)$, $q(\cdot)$, and $u(\cdot)$ are some functions. The conjugate prior to the likelihood function in (44) is

$$f_{\Theta}(\theta|\eta_0,\lambda_0) = p(\eta_0,\lambda_0) q(\theta)^{\lambda_0} e^{\lambda_0 \theta^T \eta_0},$$

(45)

where $p(\cdot,\cdot)$ is a function of $\eta_0$ and $\lambda_0$. Combining (44) and (45), the posterior distribution, given a set of observations $X = \{x_1,\ldots,x_N\}$, can be written as

$$f_{\Theta}(\theta|X;\eta_N,\lambda_N) \propto q(\theta)^{\lambda_0 + N} e^{\theta^T (\sum_{n=1}^N u(x_n) + \lambda_0 \eta_0)},$$

(46)

The posterior distribution in (46) has the same mathematical form as the likelihood function in (44), up to a normalization factor. For any distribution belongs to the exponential family, it is straightforward to obtain the posterior distribution. The closed-form solution for conjugate pairs facilitates Bayesian estimation for distributions belong to the exponential family.

The posterior distributions of the parameters in the Gaussian distribution has been introduced in [8]. For non-Gaussian distributions, we studied the conjugate prior of the beta distribution in [28] (i.e., paper A). Since the beta distribution is a member of the exponential family, it has a conjugate prior. This prior distribution is

$$f_{U,V}(u,v|\alpha_0,\beta_0,\nu_0) \propto \left[ \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} \right]^{\nu_0} e^{-\alpha_0(u-1)-\beta_0(v-1)},$$

(47)

where $\alpha_0$, $\beta_0$, and $\nu_0$ are the prior hyperparameters. Combining the likelihood function in (17) and the prior distribution in (47), the posterior distribution can be obtained as

$$f_{U,V}(u,v|X;\alpha_N,\beta_N,\nu_N) \propto \left[ \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} \right]^{\nu_N} e^{-\alpha_N(u-1)-\beta_N(v-1)},$$

(48)

where $\nu_N = \nu_0 + N$, $\alpha_N = \alpha_0 - \sum_{n=1}^N \ln x_n$, and $\beta_N = \beta_0 - \sum_{n=1}^N \ln(1 - x_n)$ are the posterior hyperparameters.
Unfortunately, this posterior is difficult to use in practice since it leads to analytically intractable integrals. There are several ways to handle this problem. Using Gibbs sampling [27], the posterior mean can be obtained by taking the mean of a set of samples generated according to the posterior distribution. By approximate inference methods [8,28,56], the posterior distribution can be approximated by another distribution, subject to some cost functions (e.g., to minimize the Kullback-Leibler (KL) divergence).

**Approximate Inference**

The central task in Bayesian estimation is to obtain the posterior distribution of the parameter variable $\Theta$ given the observed data $X = \{x_1, \ldots, x_N\}$ and evaluate some quantity with respect to this distribution. In other words, when the mathematical expression of the posterior distribution is known, the hyperparameters need to be estimated. Besides the straightforward solution above obtained by using conjugate pairs, other stochastic approximation schemes can also be applied.

**Variational Inference**

The Variational Inference (VI) framework is a general strategy for the inference of probability distributions [8, 49, 51, 56]. Suppose that the true posterior distribution $f_{\Theta}(\theta|X)$ is not practical, then a practical approximation $g_{\Theta}(\theta)$ of the true posterior distribution might be computed as an alternative solution. The logarithm of the marginal probability can be decomposed as

$$
\ln f(X) = \int g_{\Theta}(\theta) \ln \left[ \frac{f(X, \theta)}{g_{\Theta}(\theta)} \right] d\theta + \int g_{\Theta}(\theta) \ln \left[ \frac{g_{\Theta}(\theta)}{f_{\Theta}(\theta|X)} \right] d\theta = L (g_{\Theta}) + KL (g_{\Theta} \parallel f_{\Theta}(\theta|X)) = L (g) + KL (g \parallel f). (49)
$$

The first term in the last line of (49) is a lower bound of $\ln f(X)$ since the KL divergence is nonnegative. The KL divergence vanishes when $g_{\Theta}(\theta)$ equals $f_{\Theta}(\theta|X)$. As the logarithm of the marginal probability is fixed, minimizing the KL divergence is equivalent to maximizing the lower bound $L (g)$. Thus we can obtain an optimal approximation to the posterior distribution by maximizing the lower bound. Meanwhile, the form of the approximating distribution $g_{\Theta}(\theta)$ should be chosen carefully so that the KL divergence can be minimized and $g_{\Theta}(\theta)$ is feasible to work with in practical problems.

The remaining problem in VI is how to maximize the lower bound. A common strategy is to factorize the vector variable $\Theta$ into $K$ subgroups as

$$
g_{\Theta}(\theta) = \prod_{k=1}^{K} g_{\Theta_k}(\theta_k). (50)
$$

The above decomposition assumes mutual independence among the subgroups. If we also assume for a moment that variable subgroup $l$ is the only variable, the
lower bound can be factorized as [8]

\[
L(g) = \int \left( \prod_{k=1}^{K} g_{\Theta_k}(\theta_k) \right) \left[ \ln f(X, \bar{\theta}) - \sum_{k=1}^{K} \ln g_{\Theta_k}(\theta_k) \right] d\bar{\theta} \\
= \int g_{\Theta_l}(\theta_l) \left[ \int \ln f(X, \bar{\theta}) \prod_{k=1, k \neq l}^{K} g_{\Theta_k}(\theta_k) d\theta_k \right] d\theta_l \\
- \int g_{\Theta_l}(\theta_l) \ln g_{\Theta_l}(\theta_l) d\theta_l + \text{const} \\
= \int g_{\Theta_l}(\theta_l) E_{k \neq l}[\ln f(X, \bar{\theta})] d\theta_l \\
- \int g_{\Theta_l}(\theta_l) \ln g_{\Theta_l}(\theta_l) d\theta_l + \text{const} \\
= -KL \left( g_{\Theta_l}(\theta_l) \parallel e^{E_{k \neq l}[\ln f(X, \bar{\theta})]} \right) + \text{const},
\]

where

\[
E_{k \neq l}[\ln f(X, \bar{\theta})] = \int \ln f(X, \bar{\theta}) \prod_{k=1, k \neq l}^{K} g_{\Theta_k}(\theta_k) d\theta_k
\]

denotes the expectation of \( f(X, \bar{\theta}) \) with respect to all the variable subgroups except for subgroup \( \Theta_l \). The decomposition in (51) splits the lower bound into two parts, one part containing only the variable subgroup \( \Theta_l \), and one part is a constant when the variable subgroup \( \Theta_l \) is considered as the only variable. By recognizing the first part as the negative KL divergence of \( e^{E_{k \neq l}[\ln f(X, \bar{\theta})]} \) from \( g_{\Theta_l}(\theta_l) \), the lower bound \( L(g) \) can be maximized when the negative KL divergence vanishes. This indicates an optimal solution for maximizing the lower bound, which is

\[
\ln g_{\Theta_l}(\theta_l) = E_{k \neq l}[\ln f(X, \bar{\theta})] + \text{const}. 
\]

The above solution is only optimal for variable subgroup \( \Theta_l \). The optimal solution to the entire posterior distribution can be obtained by processing through all the variable subgroups, one after the other. Since the optimization problem with respect to the subgroup \( \Theta_l \) is a convex problem [101], a unique global maximum exists. The above procedure is named as the Factorized Approximation (FA) [8], and was originally developed as mean field theory in statistical physics [102].

Another advantage of the FA is that it provides an analytically tractable solution for approximating the posterior distribution, if the mathematical form of \( \ln g_{\Theta_l}(\theta_l) \) is the same as that of \( E_{k \neq l}[\ln f(X, \bar{\theta})] \). An example for the Gaussian distribution has been introduced in [8], where the mean and the precision parameters were assumed to be independent.

However, it is not always true that the terms in (53) have the same mathematical form. Further approximations may thus be required to “force” these two items to have the same expression. To this end, we extend the lower bound by replacing \( \ln f(X, \theta_l) = E_{k \neq l}[\ln f(X, \bar{\theta})] \) with a (unnormalized) lower bound \( \ln \tilde{g}(X, \theta_l) \leq \ln f(X, \theta_l) \). This lower bound should have the same mathematical expression as \( \ln g_{\Theta_l}(\theta_l) \). With this substitution, the lower bound \( L(g) \) is lower.
bounded by [28] (i.e., paper A)

\[ L(g) = -\text{KL} \left( g_{\theta_l}(\theta_l) \| \tilde{f}(X, \theta_l) \right) + \text{const} \geq -\text{KL} \left( g_{\theta_l}(\theta_l) \| \tilde{g}(X, \theta_l) \right) + \text{const}. \]  

(54)

Similar to the FA, the optimal solution is

\[ \ln g^*_{\theta_l}(\theta_l) = \ln \tilde{g}(X, \theta_l). \]  

(55)

This Extended Factorized Approximation (EFA) procedure maximizes the lower bound of the objective function \( L(g) \) instead of the objective function itself directly. Generally speaking, maximizing this lower bound will asymptotically maximize the objective function. The approximation performance depends on the shape of the (unnormalized) distribution \( \tilde{g}(X, \theta_l) \), the sharper the better. If equality holds, e.g., the (unnormalized) distribution \( \tilde{g}(X, \theta_l) \) is identical to \( \tilde{f}(X, \theta_l) \), the EFA method is equivalent to the FA method.

In paper A, we extended FA to EFA and applied this VI strategy to factorized the parameters of the prior/posterior distribution of the beta distribution (i.e., (47) and (48)) into two independent group as

\[ f_{U,V}(u,v) \approx f_U(u) f_V(v). \]  

(56)

\( f_U(u) \) and \( f_V(v) \) were assigned gamma distributions to capture the nonnegativity of \( U \) and \( V \). Based on this EFA method and relative convexity [103], we approximated the log-inverse-beta function and the pseudo digamma function with their first and second order Taylor expansions [104, 105] and derived an analytically tractable solution. Fig. 6 shows a comparison between the true posterior distribution \( f_{U,V}(u,v) \) and the variational product approximation \( f_U(u) f_V(v) \). Evidently, factorization cannot capture the correlation between \( U \) and \( V \). It is a concession for speeding up the estimation procedure by an analytically tractable solution. The accuracy of the approximation improves as the number of observations increases. More details are available in the attached paper A.

Also, for data with bounded support, we modeled the distribution of the data with a beta PDF, and assigned a gamma prior to the parameters in the beta distribution. By taking advantage of the Low Rank Matrix Approximation (LRMA), the two parameter matrices were factorized by the NMF strategy. The EFA method was applied to derive a closed-form solution for parameter estimation. More details about this beta-gamma NMF (BG-NMF) can be found in the attached paper F.

In the VI framework, if we factorize the variables into subgroups but cannot find a closed-form solution for updating each variable subgroup as in (53) or (55), we could also maximize the lower bound \( L(g) \) directly by taking the derivative with respect to each variable subgroup one by one. An example of such a solution can be found in [32].

Because of the factorization and lower bound approximation, both the FA and the EFA introduce a systematic gap between the true optimal solution and the approximate solution [6]. However, as the VI framework can provide a deterministic solution and facilitate estimation compared to sampling methods, it is still an efficient and easily implemented method for Bayesian parameter estimation.
Figure 6: Comparisons of the true posterior distribution and the approximation obtained with the VI based method. The data was generated from a Beta(x; 5, 8) distribution. The average Euclidean distances between the obtained posterior mean and the true parameters are 5.73 and 1.09 while the systematic biases of the posterior mean estimation (measured in Euclidean distance) are 5.23 and 0.32 for $N = 10$ and 100, respectively. This figure is copied from [28], i.e., the attached paper A.

**Expectation Propagation** The VI framework based method mentioned above approximates the posterior distribution $f_{\Theta}(\theta|X)$ with $g_{\Theta}(\theta)$ by minimizing the KL divergence of $f_{\Theta}(\theta|X)$ from $g_{\Theta}(\theta)$ as

$$g_{\Theta}(\theta) = \arg\min_{g_{\Theta}(\theta)} \text{KL}(g_{\Theta}(\theta) \parallel f_{\Theta}(\theta|X)).$$

(57)

In this section, we introduce and discuss the Expectation Propagation (EP) method [8, 106, 107], which is another form of deterministic approximate inference for approximating the posterior distribution in Bayesian analysis. The EP method is an improved version of Assumed Density Filtering (ADF) [108,109], so that the result obtained from EP does not depend on the ordering of the input sequence. In EP, the posterior distribution is considered to be a product of a set of factor distributions as

$$f_{\Theta}(\theta|X) \propto f_0(\theta) \prod_{n=1}^{N} f_n(\theta),$$

(58)

where $f_0(\theta)$ denotes the prior distribution and $f_n(\theta) = f_X(x_n|\theta)$ is the likelihood function of $\theta$ given the $n$th observation $x_n$. Furthermore, EP approximates the posterior distribution by a distribution $g_{\Theta}(\theta)$, which is assumed to be a product of distribution factors as

$$g_{\Theta}(\theta|X) \propto \prod_{n=0}^{N} g_n(\theta),$$

(59)
where $g_n(\theta)$ is an approximate PDF of $\Theta$ with different hyperparameters for every $n$, i.e., $g_n(\theta) = g_\Theta(\theta; \omega_n)$. In contrast to VI, the EP method minimizes the KL divergence of $g_n(\theta|X)$ from $f_n(\theta|X)$ as

$$g_n(\theta) = \arg\min_{g_n(\theta)} \text{KL}(f_n(\theta|X) \| g_n(\theta)).$$

The EP method computes an approximation by optimizing each factor $g_k(\theta), k = 1, \ldots, N$ in turn with respect to all the other factors, to ensure that the approximating distribution $g_n(\theta|X) \propto \prod_{n=0}^N g_n(\theta)$ is as close as possible to $f_n(\theta) \prod_{n \neq k} g_n(\theta)$. In other words, we refine one factor $g_k(\theta)$ while holding the remaining factors fixed. The hyperparameters of $g_k(\theta)$ are obtained based on a combination of the $k$th observation $f_k(\theta)$ and the remaining factors. It has been shown [8] that the optimal solution to the problem stated above corresponds to matching the expected sufficient statistics. In each iteration, we select one factor and update it by moment matching. This is repeated until convergence. If both the likelihood function and the prior/posterior factor distribution are Gaussian, an analytically tractable solution is obtained. Otherwise, we can apply sampling method, e.g., the importance sampling [8], to generate samples and calculate sufficient statistics for moment matching.

In [110], i.e., the attached paper C, we approximated the posterior distribution
of the parameters in the beta distribution by a product of Gaussian distributions using the EP framework. This EP based method is different from the VI based method proposed in the attached paper A. The differences are illustrated in Fig. 7. The EP based method applied a Gaussian approximation (with boundary truncation) to capture the correlation between two parameters \( u \) and \( v \) in (48), while the VI based method violated this correlation. In the moment matching step of EP, importance sampling method was utilized, while the VI based method provided an analytically tractable solution. When the amount of data is small, the EP based method outperforms the VI based method, since it contains all the information contributed from every observation by iteratively updating one factor once a time. As the amount of observations increases, both EP and VI lead to good approximations, especially for point estimate. More details about the EP based method and the comparison can be found in the attached paper C.

**Predictive Distribution**

In some practical problems, it is more interesting to study the predictive distribution. A simple way to obtain the predictive distribution of \( x \) given previous observed data \( X = \{x_1, \ldots, x_N\} \) is to plug in a point estimate of the parameters as

\[
f_X (x|X) \approx f_X(x; \hat{\theta}),
\]

where \( \hat{\theta} \) is a point estimate obtained from, e.g., ML estimation, based on the observations \( X \). However, in Bayesian analysis, we compute the entire distribution of the parameter vector. Thus a single point estimate may not be sufficient to describe all statistical characteristics. Taking the uncertainty of the parameters into account, the predictive distribution can be formulated, following a standard Bayesian approach, as

\[
f_X (x|X) = \mathbb{E}_{f_\Theta(X)} [f_X(x|\theta)],
\]

where \( f_\Theta(X) \) is the posterior distribution obtained based on \( X \). If the likelihood function \( f_X(x|\theta) \) is Gaussian and we assign an inverse-gamma distribution as the prior distribution for the variance, the predictive distribution becomes Student's \( t \)-distribution [8]. However, in most cases the predictive distribution does not have an analytically tractable solution. In such situations, a numerical method can be used to simulate the predictive distribution accurately, or some approximations can be applied to derive an analytically tractable approximation of the predictive distribution.

**Sampling Methods** Sampling methods [8, 52, 53] are numerical solution to simulate a target distribution by generating a sufficiently large number of independent samples from it. For the purpose of obtaining the predictive distribution, the distribution defined in (62) can be reformulated as

\[
f_X (x|X) = \mathbb{E}_{f_\Theta(X)} [f_X(x|\theta)],
\]

which is the expectation of \( f_X(x|\theta) \) with respect to \( f_\Theta(X) \). According to the posterior distribution, we can generate a set of independent samples \( \{\theta_1, \ldots, \theta_L\} \).
after which the expectation in (63) can be approximated by a summation as

$$ f_X (x | X) \approx \frac{1}{L} \sum_{l=1}^{L} f_X (x; \theta_l). \tag{64} $$

Obviously, the more samples we have from the posterior distribution, the more accurate the approximation in (64) would be. However, it is not always feasible to sample from the posterior distribution directly. In that case, some basic sampling methods [8], e.g., importance sampling, or rejection sampling, can be applied to generate samples according to the posterior distribution, based on a reference distribution which is easy to sample from. The performance of such kind of sampling methods depends on how well the reference distribution matches the target distribution. Moreover, methods based on Markov Chain Monte Carlo (MCMC) techniques [52, 53], e.g., the Gibbs sampling [54], are widely used as an alternative to the basic sampling methods, especially when sampling from a high-dimensional space.

**Local Variational Inference** An analytically tractable solution to the predictive distribution is preferred as it could facilitate calculations. To this end, some approximation can be applied for efficiently calculating the predictive distribution. Different from the “global” Variational Inference (VI) based method introduced in section 3.2, the Local Variational Inference (LVI) method, which is also from the variational framework, can be applied as an alternative way to compute an approximation of the predictive distribution, instead of simulating it. Recall the expression of the predictive distribution in (62), if there exists a function $g (\theta, \gamma)$ so that $g (\theta, \gamma) \geq f_{\Theta} (\theta | X)$, then an upper bound of the predictive distribution can be obtained as

$$ f_X (x | X) \leq \int_{\theta \in \Omega_{\Theta}} f_X (x | \theta) g (\theta, \gamma) d\theta \triangleq J (\gamma, x), \tag{65} $$

where $\gamma$ is a free parameter vector. If the integration in (65) can be calculated analytically, then the upper bound of the predictive distribution is a function of only $\gamma$ ($x$ is treated as a known value in this case). Minimizing the upper bound of the predictive distribution and then normalizing it afterwards, we can obtain an approximation of the value of the true predictive PDF as

$$ f_X (x | X) \approx \frac{1}{C} \min_{\gamma} J (\gamma, x) = \frac{J (\gamma^*, x)}{C}, \tag{66} $$

where $C$ is the normalization factor. Thus the remaining problem is how to find the optimal $\gamma^*$ so that the upper bound is minimized. It can be obtained by solving an optimization problem [101], subject to certain constraints, as

$$ \min_{\gamma} J (\gamma, x) \tag{67} $$

s.t. constraints.

---

3 $\gamma$ is a parameter vector in a sample space $\Omega_{\gamma}$, in which all the possible values of $\gamma$ satisfy this inequality.
To be noted, the optimal value $\gamma^*$ of (67) depends on $x$. To facilitate the calculation of the normalization factor $C$, we can take a reasonable approximation to $\gamma^*$, regardless of the value of $x$ [111]. In general, the optimized upper bound $J(\gamma^*, x)$ is not exactly the same as the first term in (65), thus the result is not accurate. However, it provides a convenient way to approximate the integration with an analytically tractable solution.

In [111], i.e., the attached paper B, we replaced the posterior distribution of the parameters in the beta distribution, as obtained from the attached paper A, by an upper bound. With some mathematical analysis, we proved the existence of a global minimum of the upper bound. The predictive distribution of the beta distribution is then approximated by an analytically tractable solution, which was shown to be more similar to the true one, compared to the predictive distribution approximated by plugging in the point estimate of the parameters. More details about this LVI based method for approximating the predictive distribution of the beta distribution can be found in the attached paper B.

3.3 Model Selection

When describing data using statistical model, considerable attention should be paid to which model, from a set of possible descriptions, is to be preferred, in addition to parameter estimation. Several criteria have been proposed and applied for the purpose of model selection. Frequently used criteria, among others, include the Akaike Information Criterion (AIC) [112], the Bayesian Information Criterion (BIC) [98], and the Bayes Factor (BF) [113]. The AIC is based on the concept of information theory, which is used to measure the loss of the KL divergence of the hypothesis model from the true one, so the smaller the better. The AIC is defined as

$$\text{AIC} = -2 \ln f(\hat{X}|\hat{\theta}) + 2s,$$

where $X = \{x_1, \ldots, x_N\}$ is the observed data and $s$ denotes the degree of freedom of parameters in the statistical model.

For the purpose of Bayesian model selection, the BIC, also called the Schwarz criterion, was proposed in order to favor simpler models than those chosen by the AIC. The BIC applies more weight to the model size and penalizes it as

$$\text{BIC} = -2 \ln f(\hat{X}|\hat{\theta}) + s \ln N.$$

The model with a smaller BIC value is preferred. A comparison of AIC and BIC is available in [114].

The BF offers a way to evaluate evidence in favor of a null hypothesis [113]. It is a more general criterion for Bayesian model comparison. In the Bayesian framework, the BF of model hypothesis $H_1$ over $H_2$ is

$$\text{BF} = \frac{f(X|H_1)}{f(X|H_2)}$$

where

$$f(X|H_k) = \int_{\theta_k \in \Omega_{\theta_k}} f(X|\theta_k) f(\theta_k|H_k) d\theta_k, \quad k = 1, 2$$
represents the model evidence for the observations X under hypothesis \( H_k \), and \( \theta_k \) is the parameter vector under that hypothesis. If the prior probabilities of the two hypotheses are equal, the BF is equal to the ratio of the posterior hypothesis probabilities given the data, since

\[
\frac{p(H_1|X)}{p(H_2|X)} = \text{BF} \times \frac{p(H_1)}{p(H_2)}. \tag{72}
\]

Actually, the BIC is an approximation of the logarithm of the BF. As the BF is difficult to calculate in general, the BIC is usually used as an alternative. When the number of parameters are equal in both statistical model hypotheses, the BF and the BIC give identical results.

In our research, when comparing the performance of non-Gaussian statistical models to Gaussian statistical model, or when deciding the complexity (the number of free parameters) of a mixture model, we utilized both the BIC and the BF.

### 3.4 Graphical Models

In statistical models, the relations among the variables, no matter how complex they are, are usually represented by mathematical expressions. Graphical models provide a way to visualize the structure of statistical models [49, 56, 115]. A graphical model contains nodes, the variables, and edges, the probabilistic relations among different variables. For Bayesian analysis, Directed Acyclic Graph (DAG) is used to infer the (conditional) dependencies and independencies among the variables. DAG is also called a Bayesian network [8]. Fig. 8 illustrates the conditional dependence/independence of two variables, also known as the Bayes ball algorithm [115]. A Bayesian network consists of several basic probabilistic relations of the kind is described in Fig. 8. According to that, the statistical model can be separated into several groups to facilitate the analysis. Sometimes, (conditional) independence is introduced to simplify the problem, by modifying the connections in Bayesian network. We utilized Bayesian network for analyzing the probabilistic relations among the variables in the attached papers A and F.

### 4 Applications of Non-Gaussian Models

For data with bounded support or semi-bounded support, it has been shown that non-Gaussian statistical models and the corresponding non-Gaussian Mixture Model (nGMM) can provide better modeling performance than Gaussian models [27, 28, 30–32, 74, 87, 89] according to the criteria introduced in section 3.3. We thus expect these better models can lead to improved performance for different aspects in practical applications. In our research, we investigated the use of non-Gaussian statistical models for several applications.

#### 4.1 Speech Processing

Speech processing is an application of digital signal processing to speech. This includes speech coding, speech recognition, speaker recognition, to mention a few.
Speech signals are not stationary processes, so the Linear Prediction (LP) method is applied to a short segment of speech (usually 20∼30 ms block length and modified by a window function) to estimate a linear representation of the speech signal in each frame [57]. The LP model with prediction order $K$ is formulated as

$$\hat{x}(t) = \sum_{k=1}^{K} \rho_k x(t - k),$$

where $x(t)$ is the speech sample at time $t$, and $\rho_k$ denotes the Linear Prediction Coefficient (LPC). The parameters in the LP model can be estimated in several ways, e.g., the autocorrelation method [117, 118], the covariance matrix
method [57]. Then the LP analysis filter can be expressed by z-transform as

\[ G(z) = 1 - \sum_{k=1}^{K} \rho_k z^{-k}, \]  \hspace{1cm} (74)\]

which is known as the “whitening” filter as it removes the short-term correlation and flattens the spectrum [119]. As the LP filter is a minimum-phase filter, the synthesis filter, which is the inverse of the LP analysis filter, exists and is also stable. Fig. 9 shows an example of speech and the corresponding envelope obtained by LP synthesis filter.

In speech coding, an essential part is quantizing the LPCs efficiently. Directly quantizing the LPCs is usually not used because small quantization errors
in the LPCs may result in relative large spectral distortion and could lead to the instability of the filter [119]. Thus, some representations of the LPCs have been proposed for the purpose of effective quantization. The most used representations are the Reflection Coefficient (RC), the ArcSine Reflection Coefficient (ASRC), the Log Area Ratio (LAR), the Line Spectral Frequency (LSF), and the Impedance Spectral Frequency (ISF) [57, 119–121]. Among these representations, the LSF is the widely used representation when quantizing the LP model. With the analysis filter in (74), two symmetric polynomials can be created as [119]

$$P(z) = G(z) + z^{-(K+1)}G(z^{-1}),$$

$$Q(z) = G(z) - z^{-(K+1)}G(z^{-1}).$$

(75)

The zeros of $P(z)$ and $Q(z)$ are interleaved on the unit circle as (by assuming that $K$ is even)

$$0 = \phi_{q0} < \phi_{p1} < \phi_{q1} < \ldots < \phi_{qK/2} < \phi_{pK/2 + 1} = \pi.$$  

(76)

Then LSF parameters are defined as [120]

$$\mathbf{s} = [s_1, s_2, \ldots, s_K]^T = [\phi_{p1}, \phi_{q1}, \ldots, \phi_{pK/2}, \phi_{qK/2}]^T.$$  

(77)

The modification effect of the LSF parameters is local, which means that modifying the value of one element in the LSF vector will only locally affect the envelope in a small area. Fig. 9(d) illustrates this local effect of modifying the LSF parameter. As mentioned above, the LSF parameters are ordered and bounded in $[0, \pi]$. To exploit these properties, several quantization methods of the LSF parameters were studied in the literature (see e.g., [122–125]). The problem of designing an optimal quantization scheme for the LPC parameters has also been investigated by many researchers [73, 123, 126]. The well-known Lloyd [127] and the Linde-Buzo-Gray (LGB) algorithms [128] are usually utilized for obtaining the codebook. However, these algorithms depend on the training data and lead to worse quantization performance if the training data is not representative or the amount of training data is limited.

Recently, the PDF-optimized Vector Quantization (VQ) scheme was proposed [73, 129, 130, 130] to overcome the problem of lacking training data. For this purpose, the GMM is widely applied to model the underlying distribution of the LSF parameters. Based on the trained model, a training set with sufficient amount of data (theoretically infinite) can be generated and used for obtaining the codebook. Also, to prevent training a VQ in a high-dimensional space, the transform coding [131] technique is often applied to decorrelate the vector variable into a set of independent scalar variables. Thus, the VQ can be replaced by a set of Scalar Quantization (SQ) without losing the memory advantage of VQ. For Gaussian source, the Karhunen-Loève Transform (KLT) is usually applied. By the high rate theory in source coding [58], the optimal bit allocation strategy can be derived based on the trained model and an entropy coding scheme can also be derived afterwards [130].

As suggested in [129], explicitly taking the bounded support constraint of the LSF parameters into account could improve the VQ performance. Based on
this, Lindblom et al. [29] proposed a bounded support GMM based VQ, which is a modified and improved version of the conventional GMM based VQ, both in model estimation and in VQ design. However, the estimation algorithm of the bounded support GMM is computationally costly when the bounded support is taken into account [29]. Also, a lot of mixture components were spent on describing the edge of the distribution. To avoid the extra computational cost and save the bit budget, we applied the BMM based VQ for the quantization of the LSF parameters by only considering the bounded support property. The performance of the BMM based VQ was shown to be superior to the GMM based VQ with the same level of model complexity (in terms of the number of parameters in the statistical model). We introduced this work in [87].

To further exploit the ordered property, we transformed the LSF parameters linearly to another representation, the LSF difference (∆LSF), as

\[ \bar{x} = \varphi(x) = A \omega, \quad A = \frac{1}{\pi} \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & -1 \end{bmatrix}_{K \times K}. \]  

(78)

The linear transformation in (78) is identically invertible. Since there is no information loss during the transformation, quantizing the ∆LSF parameters is equivalent to quantizing the LSF parameters. The ∆LSF vector \( \bar{x} \) contains positive elements and the sum of all the elements is smaller than \( \pi \). Thus \( \bar{x} \) has the same domain as the Dirichlet variable, up to a normalization constant \( \pi \). Hence, we applied a Dirichlet Mixture Model (DMM) to model the underlying distribution of the ∆LSF parameters, by considering both the boundary and the ordered properties of the LSF parameters. As the Dirichlet variable is neutral [81], we proposed a non-linear transformation to decorrelate the Dirichlet vector variable into a set of independent beta variables, which is similar as the KLT for the Gaussian variables but not linear. After that, an optimal bit allocation strategy was derived based on the distortion-rate (D-R) relation. Finally, a practical coding scheme similar to the Differential Pulse Code Modulation (DPCM)\(^4\) was proposed for preventing the error propagation in the sequential quantizations. The modeling performance of the ∆LSF parameters was studied in the attached paper E. The details of the DMM based VQ design and implementation can be found in the attached paper D.

Another important application in speech processing is speaker recognition [132], which includes two tasks: 1) Speaker Identification (SI), to identify a particular speaker [133]; and 2) Speaker Verification (SV), to verify a speaker’s claimed identity [134, 135]. For both tasks, the Mel-Frequency Cepstral Coefficient (MFCC) [59] is the widely used feature to represent the speaker’s characteristics. For the task of SI, the LSF parameter is also used as a feature presentation [132, 136], since the LP filter parameters are determined by the speaker’s articulatory system, and the LSF parameters convey information about the speaker’s

\(^4\)A similar approach was also used in [124], in which a empirical grouping strategy was applied to the LSF parameters.
identity.

The MFCC feature suppresses the spectrum details in the high frequency area by introducing a mel-scale filter bank, which takes the advantages of human ears’ frequency selectivity. However, the information contained in the high frequency area might be useful for the machine to identify the speaker. In other words, the LSF contains “full band” information and might perform better than the band-modified MFCCs. Hence, we used the ∆LSF parameters, which contains the same information as the LSF parameters, as the feature for the task of SI [137]. Since the dynamic information is useful for speaker recognition, we represented the dynamic information of the ∆LSF parameters by considering two neighbors of the current frame, one from the past frames and the other from the following frames. These ∆LSF vectors are cascaded to make a supervector. To model the distribution of the supervectors, the so-called super-Dirichlet distribution and the corresponding super-Dirichlet Mixture Model (sDMM) was proposed. The sDMM based SI system was shown to be superior to the GMM based system, for the task of text-independent speaker identification [137].

4.2 Image Processing

For digitalized images, the pixel-histogram based method is a simple but efficient way for the purpose of image classification [138], image segmentation [139, 140], etc. In most cases, the Gaussian distribution and the corresponding GMM are used to describe the distribution of pixel values [23, 139–142]. Actually, the pixel values are located in a fixed interval \([0, 2^R - 1]\), where \(R\) is the number of bits for storing a pixel in the computer. Thus the assumption of Gaussian distribution of the pixel is not accurate, as it violates the boundary property [25, 27, 79, 84].

To model the distribution of pixel values efficiently, Bouguila et al. [27] modeled the distribution of the gray image pixels by a Beta Mixture Model (BMM) and proposed a practical Gibbs sampling method for parameter estimation. We applied the BMM to describe the histogram of the handwritten digits for the purpose of classification [79]. The color image is usually described and stored in the RGB space. One pixel in the color image is a composition of three channels, each of which has a value in a fixed interval \([0, 2^R - 1]\). The Dirichlet distribution and the corresponding DMM were used to model the distribution of the color pixels in the RGB space and showed an improvement over the GMM based method [75, 84]. In these papers, the human skin color pixel in the RGB space was firstly normalized by the sum of three channels, then the distribution of the normalized pixel vectors was modeled by a DMM. This normalization could remove the illuminance so that the variance of each channel can be reduced [143]. However, the variance of the non-skin color pixel may also be reduced as well. Furthermore, if the illuminance of the color pixel is removed, the color pixels containing the different illuminances but the same RGB proportions will be recognized as the same cluster. For example, the color “light green” [0, 255, 0] and “deep green” [0, 100, 0] have the same proportion [0, 1, 0] and may not be distinguished efficiently. As mentioned in page 12, the domain of the three-dimensional

\footnote{When taking the MFCC as the feature, the ∆MFCC and the the ∆∆MFCC are usually combined with the MFCC to describe the dynamic information.}
Dirichlet variable is a simplex while the domain of a three-dimensional beta variable is a unit cubic. According to the above discussion, we considered the color pixel as an extension of the gray image pixel and applied the multivariate beta distribution [28,144] to model the color pixels’ distribution for the task of human skin color detection [144]. The BMM based method outperformed some other methods based on the pixel probabilistic model. In general, the non-Gaussian statistical model based methods are better than the Gaussian statistical model based method, in terms of the ability of describing the pixel histogram [27,28,75].

4.3 Nonnegative Matrix Factorization

The Nonnegative Matrix Factorization (NMF) is an important technique for matrix decomposition. It decomposes a nonnegative matrix into a product of two nonnegative matrices [145,146]. The NMF can be applied to several applications, such as face feature extraction [146], image denoising [24], sparse coding [147], music spectrum modeling [32,148], etc. The conventional form of NMF can be expressed as

$$X_{P \times T} \approx WP_{P \times K}V_{K \times T},$$

(79)

where $x_{pt}$, $w_{pk}$, and $v_{kt}$ are nonnegative elements and $p = 1, \ldots, P$, $t = 1, \ldots, T$, $k = 1, \ldots, K$. If we choose $K < \min (P, T)$, then the NMF is a Low Rank Matrix Approximation (LRMA). By the expression in (79), a column $\overline{x_t}$ in $X$ can be interpreted as a linear combination of the columns in $W$, weighted by the factors in the $t$th column of $V$, as

$$\overline{x_t} = \sum_{k=1}^{K} w_{tk} v_{kt}.$$  

(80)

As the observation $X$ contains nonnegative data (e.g., image data, speech spectra) in various applications, the NMF can serve a better performance for the nonnegative data [146], compared to some other classical LRMA methods, e.g., Singular Value Decomposition (SVD).

The conventional way to is to minimize the distance between $X_{P \times T}$ and $WP_{P \times K}V_{K \times T}$ as

$$\min_{W,V} D (X||WV)$$

s.t. $w_{pk} \geq 0$, $v_{kt} \geq 0,$

(81)

where $D (\cdot || \cdot)$ denotes a distance measure. For example, by taking the Frobenius norm for measuring the distortion, the NMF algorithm tries to minimize the Euclidean distance between the original matrix and the approximated one as [149]

$$\min_{W,V} \|X - WV\|_F^2$$

s.t. $w_{pk} \geq 0$, $v_{kt} \geq 0.$

(82)

If the generalized KL divergence is taken into account as the criterion [149], the NMF algorithm is to solve a constrained optimization problem as

$$\min_{W,V} GKL (X||WV)$$

s.t. $w_{pk} \geq 0$, $v_{kt} \geq 0,$

(83)
where
\[
\text{GKL} (A \parallel B) = \sum_{ij} \left( a_{ij} \log \frac{a_{ij}}{b_{ij}} - a_{ij} + b_{ij} \right).
\]
(84)

This measure reduces to the standard KL divergence when \( \sum_{ij} a_{ij} = \sum_{ij} b_{ij} = 1 \).

Another distortion measure for NMF is the Itakura-Saito (IS) distance as:
\[
\min_{W, V} \text{IS} (X \parallel WV)
\text{ s.t. } w_{pk} \geq 0, v_{kt} \geq 0
\]
(85)

with
\[
\text{IS} (A \parallel B) = \sum_{ij} \left( \frac{a_{ij}}{b_{ij}} - \log \frac{a_{ij}}{b_{ij}} - 1 \right).
\]
(86)

The IS-NMF showed a promising improvement over the above two methods when modeling music spectra [148]. Generally speaking, the optimization problem connected to the NMF algorithm is convex with respect to \( W \) or \( V \). But it is not jointly convex in terms of \( W \) and \( V \). Thus, the gradient methods are usually utilized to solve this constrained optimization problem approximately [147–149].

The above mentioned NMF frameworks, based on different criteria, can all be interpreted in a probabilistic way in terms of different probability distributions [148]. To prevent the overfitting problem, the NMF can also be carried out in a Bayesian framework. The minimization problem in (81) can be expressed by a maximization problem with respect to an appropriate PDF as
\[
\{W^*, V^*\} = \arg \max_{W, V} f (X; WV, \Theta),
\]
(87)

where \( \Theta \) denotes the (possible) parameter set according to the choice of PDF. For example, the Euclidean distance based NMF (E-NMF) method is equivalent to the ML estimation of the parameters by assigning a Gaussian distribution for the observation as
\[
\{W^*, V^*\} = \arg \max_{w_{pk}, v_{kt}} \prod_{p,t} \mathcal{N} \left( x_{pt}; \sum_{k} w_{pk} v_{kt}, \sigma \right).
\]
(88)

Schmidt et al. [24,150] proposed two Bayesian NMF methods based on this Gaussian assumption. In [150], the prior distribution of \( w_{pk} \) and \( v_{kt} \) was assigned with a Gaussian prior, via a link function to preserve the nonnegativity. In [24], the exponential distribution was assigned to \( w_{pk} \) and \( v_{kt} \), respectively. The Gibbs sampler was used to simulate the posterior distribution. The proposed Gaussian-NMF method was applied to image feature extraction and showed a promising improvement.

The generalized KL (GKL-NMF) can be connected to the Poisson distribution as
\[
\{W^*, V^*\} = \arg \max_{w_{pk}, v_{kt}} \prod_{p,t} \mathcal{P} \left( x_{pt}; \sum_{k} w_{pk} v_{kt} \right),
\]
(89)

where
\[
\mathcal{P} (x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \text{ } x \text{ is a positive integer.}
\]
(90)
Cemgil [31] took this assumption and proposed a Bayesian inference method in NMF, by assigning gamma distributions for $w_{pk}$ and $v_{kt}$. Since the gamma distribution is the conjugate prior to the Poisson distribution and the summation of several Poisson variables is still Poisson distributed, an analytically tractable solution was obtained by the Variational Inference (VI) framework. However, the proposed Poisson-NMF method could only be suitable to variables $X_{pt}$ with discrete value. The application of the Poisson-NMF is mainly on image data [31], which is usually considered as a continuous variable with bounded support. This mismatch impairs the statistical interpretation of the GKL-NMF on non-countable data [148].

The IS-NMF is equivalent to the ML estimation of the parameters in gamma multiplicative noise with mean 1 [148], up to a scale factor and a constant. To this end, the IS-NMF method can be interpreted as [148, eq. 31]

$$\{W^*, V^*\} = \arg \max_{w_{pk}, v_{kt}} \prod_{p,t} \frac{1}{\sum_k w_{pk} v_{kt}} \text{Gam} \left( \frac{x_{pt}}{\sum_k w_{pk} v_{kt}} ; \alpha, \alpha \right),$$

(91)

where $\alpha$ is a free parameter that controls the variance of the gamma noise. A Bayesian NMF with the exponential assumption (which is a special case of the gamma assumption in [148]) was proposed in [32]. The prior distribution of $w_{pk}$ and $v_{kt}$ was assumed to be gamma. The VI framework was applied and the approximations to the posterior distributions were derived analytically. This approach was proposed mainly for the analysis of music spectra [32].

As mentioned above, when applying the Poisson-NMF to image data, it violates the continuous and bounded support properties of the pixel values. Since it has been shown [27, 28] that the beta distribution can model the pixel data better than some other distributions because of its bounded support, we proposed a beta distribution based NMF strategy for modeling the bounded support data in [33] (i.e., the attached paper F). For a nonnegative matrix $X_{P \times T}$, each element variable $X_{pt}$ has a bounded support and is assumed to be beta distributed with parameter $m_{pt}$ and $n_{pt}$ as

$$X_{pt} \sim \text{Beta}(x_{pt}; m_{pt}, n_{pt}).$$

(92)

Then for the whole matrix, we obtain two parameter matrices $M_{P \times T}$ and $N_{P \times T}$. To create a hierarchical Bayesian framework, we assume these two parameter matrices are latent random variables. If factorize these two variable matrices jointly as

$$M_{P \times T} \approx A_{P \times K} H_{K \times T},$$
$$N_{P \times T} \approx B_{P \times K} H_{K \times T},$$

(93)

and assign a gamma distribution to each element in $A_{P \times K}$, $B_{P \times K}$, and $H_{K \times T}$, a generative model for the bounded support variable can be obtained as

$$A_{pk} \sim \text{Gam}(a_{pk}; \mu_0, \alpha_0),$$
$$B_{pk} \sim \text{Gam}(b_{pk}; \nu_0, \beta_0),$$
$$H_{kt} \sim \text{Gam}(h_{kt}; \rho_0, \zeta_0),$$
$$X_{pt} \sim \text{Beta}(x_{pt} | \sum_k a_{pk} h_{kt}, \sum_k b_{pk} h_{kt}).$$

(94)
This generative model assumes the mutual independency of the elements in \( M_{P \times T} \) and \( N_{P \times T} \). However, the correlations of the elements are captured by the joint NMF in (93) since they share the same weighting matrix \( H_{K \times T} \). With the VI framework, an analytically tractable solution was derived to approximate the posterior distributions of \( A_{P \times K} \), \( B_{P \times K} \), and \( H_{K \times T} \) in [33] (i.e., the attached paper F). By definition, we have

\[
E[X_{pt} | A_{P \times K}, B_{P \times K}, H_{K \times T}] = \frac{\sum_k a_{pk} h_{kt}}{\sum_k a_{pk} h_{kt} + \sum_k b_{pk} h_{kt}}
\]

If we take the posterior mean as the point estimate to the latent variables \( a_{pk}, b_{pk}, \) and \( h_{kt} \), the expected value of \( X_{pt} \) can be approximated as

\[
E[X_{pt} | A_{P \times K}, B_{P \times K}, H_{K \times T}] \approx \frac{\sum_k a_{pk} h_{kt}}{\sum_k a_{pk} h_{kt} + \sum_k b_{pk} h_{kt}}.
\]

We proposed the beta-gamma-NMF (BG-NMF) in [33] and applied it to image processing and collaborative filtering problems. Compared to recently proposed methods [31, 147, 151, 152], it showed a promising improvement.

This proposed BG-NMF can also be interpreted by the IS-NMF, under certain condition. Recall (91), if we set \( \alpha = \sum_k w_{pk} v_{kt} \), then we have

\[
\{W^*, V^*\} = \arg \max_{w_{pk}, v_{kt}} \prod_{p,t} \text{Gam} \left( x_{pt}; \sum_k w_{pk} v_{kt}, 1 \right),
\]

which means \( x_{pt} \) is gamma distributed with shape parameter \( \sum_k w_{pk} v_{kt} \) and scale parameter 1. Furthermore, if we assume that \( Y_{pt} \) and \( Z_{pt} \) are gamma distributed as

\[
Y_{pt} \sim \text{Gam} \left( y_{pt}; \tilde{y}_{pt}, 1 \right) \quad \text{and} \quad Z_{pt} \sim \text{Gam} \left( z_{pt}; \tilde{z}_{pt}, 1 \right),
\]

then \( X_{pt} = \frac{Y_{pt}}{Y_{pt} + Z_{pt}} \) is beta distributed with parameter \( \tilde{y}_{pt} \) and \( \tilde{z}_{pt} \) as [153]

\[
X_{pt} \sim \text{Beta} \left( x_{pt}; \tilde{y}_{pt}, \tilde{z}_{pt} \right).
\]

Under the IS-NMF framework, factorizations can be placed on the two variable matrices \( Y_{P \times T} \) as

\[
Y_{P \times T} \approx \tilde{Y}_{P \times T} = A_{P \times K} H_{K \times T},
\]

\[
Z_{P \times T} \approx \tilde{Z}_{P \times T} = B_{P \times K} H_{K \times T}.
\]

Again, if we assign a gamma prior to each of the elements in \( A_{P \times K}, B_{P \times K}, \) and

\[6\text{Please note, } x_{pt} \text{ in (97) can be any nonnegative value, which is not the same as the bounded } x_{pt} \text{ in (94). The usage of } x_{pt} \text{ in (97) is for the consistence with that in (91).} \]
H_{K \times T}, a generative model can be obtained as

\begin{align*}
A_{pk} &\sim \text{Gam}(a_{pk}; \mu_0, \alpha_0), \\
B_{pk} &\sim \text{Gam}(b_{pk}; \nu_0, \beta_0), \\
H_{kt} &\sim \text{Gam}(h_{kt}; \rho_0, \zeta_0), \\
Y_{pt} &\sim \text{Gam}(y_{pt} \mid \sum_k a_{pk} h_{kt}, 1), \\
Z_{pt} &\sim \text{Gam}(z_{pt} \mid \sum_k b_{pk} h_{kt}, 1), \\
X_{pt} &= \frac{Y_{pt}}{Y_{pt} + Z_{pt}} \sim \text{Beta}(x_{pt} \mid \sum_k a_{pk} h_{kt}, \sum_k b_{pk} h_{kt}).
\end{align*}

(101)

This generative model is equivalent to that in (94). The only difference is that two intermediate variables \(Y_{pt}\) and \(Z_{pt}\) are introduced. Similarly, the expected value of the bounded variable \(X_{pt}\) is

\[E[X_{pt} \mid A_{P \times K}, B_{P \times K}, H_{K \times T}] = \frac{\sum_k a_{pk} h_{kt}}{\sum_k a_{pk} h_{kt} + \sum_k b_{pk} h_{kt}},\]

(102)

and it can also be approximated by taking the point estimate to \(A_{pk}, B_{pk},\) and \(H_{kt}\) as

\[E[X_{pt} \mid A_{P \times K}, B_{P \times K}, H_{K \times T}] \approx \frac{\sum_k A_{pk} H_{kt}}{\sum_k A_{pk} H_{kt} + \sum_k B_{pk} H_{kt}}.\]

(103)

5 Summary of Contributions

5.1 Overview

The work introduced in this dissertation focuses mainly on

- the Maximum Likelihood (ML) and Bayesian estimations of non-Gaussian statistical models,
- their applications in various fields.

This dissertation mainly consists of six papers, in which I formulated the problems and proposed the approaches to solve the problems. Also, I did the mathematical derivations and conducted the experimental evaluations. The coauthor of these papers, who is my supervisor, gave me a lot of fruitful suggestions, both on theoretical and experimental parts. These attached papers can be categorized into three packages, according to the contents covered by them.

The attached paper A, B, and C focused on Bayesian analysis of the beta distribution. By the principles of the Variational Inference (VI) framework, the posterior distribution of the correlated parameters in the beta distribution was approximated by a product of independent gamma distributions. The Factorized Approximation (FA) method was extended to Extended Factorized Approximation (EFA), by relaxing the lower bound. This approximation facilitated the estimation procedure by an analytically tractable solution. Based on this posterior
approximation, the predictive distribution can be obtained by the LVI approximately. Again, an analytically tractable expression for the predictive distribution makes the problem easy. To capture the correlation between the parameters in the beta distribution, a EP based method was proposed to approximate the posterior distribution and showed an advantage when the amount of observed data is small. In general, the beta distribution can model the bounded support data better than the conventional Gaussian distribution. The Bayesian estimation method based Beta Mixture Model (BMM) was applied for several applications and showed an improvement.

The attached paper D and E mainly focused on improving the VQ performance of the LP model for speech. By taking the boundary and ordered properties, the LSF parameters were linearly transformed to the ∆LSF parameters and a DMM was applied to model the underlying distribution. The ML estimation of the parameters in the mixture model was proposed. The Dirichlet variable was decorrelated into a set of independent beta variables, according to its aggregation property and neutrality. The optimal bit allocation strategy for inter- and intra-component was also proposed based on the estimated distribution. The proposed PDF-optimized VQ outperformed the state-of-the-art GMM based VQ for transparent coding.

In the attached paper F, we derived a beta-gamma NMF (BG-NMF) method for bounded support data. The distribution of the bounded support data matrix was modeled by a matrix of beta variables. Then the parameter matrices of the beta variables were factorized jointly, and each of the elements in the basis and excitation matrices was assigned with a gamma prior. This generative model can also be interpreted by the IS-NMF under certain condition. Based on the VI framework and the EFA method, some lower bound approximations were used to derive an analytically tractable solution. The proposed BG-NMF was applied to different applications in image processing and also used for the collaborative filtering application.

5.2 Conclusions

For data with bounded support or semi-bounded support, it has been proven that the non-Gaussian statistical models and the corresponding non-Gaussian Mixture Model (nGMM) can provide a better modeling performance [27, 28, 30–32, 74, 87, 89] by considering the criteria introduced in section 3.3. Thus, in real applications, it is also expected that a better modeling behavior can lead to a better performance for different application aspects. In our research work, we applied the non-Gaussian statistical model in several applications. Compared to some conventional statistical model based methods, the non-Gaussian model based methods show a promising improvement.

References


