Variational Inference for Watson Mixture Model

Jalil Taghia and Arne Leijon

Abstract—This paper addresses modelling data using the Watson distribution. The Watson distribution is one of the simplest distributions for analyzing axially symmetric data. This distribution has gained some attention in recent years due to its modeling capability. However, its Bayesian inference is fairly understudied due to difficulty in handling the normalization factor. Recent development of Markov chain Monte Carlo (MCMC) sampling methods can be applied for this purpose. However, these methods can be prohibitively slow for practical applications. A deterministic alternative is provided by variational methods that convert inference problems into optimization problems. In this paper, we present a variational inference for Watson mixture models. First, the variational framework is used to side-step the intractability arising from the coupling of latent states and parameters. Second, the variational free energy is further lower bounded in order to avoid intractable moment computation. The proposed approach provides a lower bound on the log marginal likelihood and retains distributional information over all parameters. Moreover, we show that it can regulate its own complexity by pruning unnecessary mixture components while avoiding over-fitting. We discuss potential applications of the modeling with Watson distributions in the problem of blind source separation, and clustering gene expression data sets.

Index Terms—Bayesian inference, variational inference, Watson distribution, mixture model, axially symmetric, clustering on the unit hypersphere, blind source separation, gene expression

1 Introduction

Directional data emerge naturally in many areas (e.g., [1, Ch. 1] for typical applications) and more recently in bioinformatics (e.g., [2]), collaborative filtering (e.g., [3]), and data mining (e.g., [4]). To analyze directional data, directional distributions are of great importance, mainly where popular distortion measures (or the corresponding generative models) based on minimizing euclidean distortions yield poor results [5]. The von Mises-Fisher (vMF) distribution is one of the fundamental distributions in analysis of directional data, which models data concentrated around a mean direction. However, data on the unit hypersphere arise in various ways depending on not only direction but also other structures. Closely related to the directional data are *axially symmetric data* for which the probability density is the same in each direction as in the opposite direction [1, Ch. 1]. Thus for modeling data with such additional structure, the limited modeling capability of the vMF distribution can be restrictive. The (Dimroth-Scheidegger-) Watson distribution [6] is one of the simplest distributions for modeling axially symmetric data [1, Ch. 9.4], and thus the focus of our paper.

1.1 Recent Applications

Beyond typical applications of the Watson distribution in shape analysis due to desirable property of rotational invariance (e.g., [7], [8]), this distribution has gained renewed attention in other application domains in recent years. Bijral et al. [9] develop a generative model of mixtures of Watson distributions for clustering spectral embeddings which often have noisy formations in the embedding space. Results of their experiments on simulated data and real data sets (USPS Zipcodes Dataset, 20 Newsgroups, and Yahoo 20) show better modeling capability of this distribution compared to the vMF distribution. For such data, the vMF model tends to put several points in the same cluster resulting in a very large estimated concentration parameter for one cluster and incorrectly small estimated concentration parameters for the other clusters, while the Watson model allows negative concentration parameters which result in more evenly distributed clusters, [9].

In another study [10], authors discuss mixture modelling with Watson distributions and reveal its connection to the “diametrical clustering” [11]. Diametrical clustering is used for identifying anti-correlated gene clusters. Many clustering algorithms cluster genes together when their expression patterns show high positive correlation. However, it has been observed that genes whose expression patterns are strongly anti-correlated can also be functionally similar [12]. Sra and Karp, [10], show that diametrical clustering can be regarded as a special case of mixture modelling with Watson distributions such that the additional modeling power can lead to better clustering (e.g., clustering gene microarray datasets).

Lennox et al. [13] proposed a Dirichlet process mixture of bivariate vMF distributions for protein configuration angles, modifying the finite mixture model of Mardia et al. [14]. Bhattacharya and Dunson [15] discuss that the vMF kernel proposed in [13] can be restrictive, and it is not clear whether mixtures of such kernels induce priors with large support. Alternatively, Bhattacharya and Dunson [15] consider complex Watson kernel and propose Dirichlet process mixtures of complex Watson distributions with applications to planar shapes. They show that such priors have large support.

A somewhat different application domain of the complex Watson mixture model is in the problem of blind source

- *The authors are with the School of Electrical Engineering, Communication Theory Lab., KTH Royal Institute of Technology, 10044 Stockholm, Sweden. E-mail: {taghia, leijon}@kth.se.*

Manuscript received 18 June 2013; revised 28 Dec. 2014; accepted 23 Oct. 2015. Date of publication 8 Nov. 2015; date of current version 11 Aug. 2016.
Recommended for acceptance by Y. Teh.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2015.2498935

separation (BSS)—a challenging task aiming in separation of a set of source signals from a set of mixed signals, with very little information about the source signals or the mixing process (e.g., [16]). For this purpose, one may interpret the values of the normalized short time Fourier transform of mixed signals to be drawn from a mixture of complex Watson distributions. Tran-Vu and Haeb-Umbach [17] employed a mixture of complex Watson distributions for clustering normalized time-frequency points in BSS of convolutive speech signals. Modeling the statistics of the data using complex Watson model has two advantages: firstly, distortion measure of the this distribution fits well to the beamforming point of view (we shall discuss this point further in Section 5.2.1). Secondly, "the morphology of the complex hypersphere perfectly reflects spatial aliasing occurring in high frequencies, hence the cyclic nature of the phase differences are implicitly regarded", [17]. In another study, Besson and Bidon [18] derived an adaptive beamformer by considering the steering vector of interest as a random variable with a prior complex Watson distribution. Authors show that this particular choice of prior can lead to significant improvement compared to the conventional approaches.

1.2 Related Work and Contributions

The main difficulty in working with the Watson distribution stems from the complicated normalizing factor involving the Kummer's confluent hypergeometric function. This makes even a simple task of maximum-likelihood (ML) parameter estimation challenging, specifically in the estimation of concentration parameters. As shown in [7], for the ML estimation of the Watson distribution, there exists no analytically tractable solution for the estimation of the concentration parameters. Numerical estimation (e.g., with Newton-Raphson methods) of the concentration parameters is nontrivial in high dimensions since it involves functional inversion of ratios of Kummer functions.

Mardia and Dryden [7] provided a simple approximation to the ML estimates, which has reasonable accuracy only for low dimensions. Bijral et al. presented a generative model of mixtures of Watson distributions and introduced *ad-hoc* approximations of the parameters in an expectation maximization (EM) setting. Sra and Karp [10] removed the ad-hoc approach by deriving tight, two-sided bounds to the ML estimates.

However, Bayesian inference of directional distributions, in particular the Watson distribution, has been fairly understudied in the literature perhaps due to difficulties in dealing with normalizing constants. In general, for Bayesian inference of the Watson distribution, one can find the prior distribution and the corresponding conjugate posterior distribution of the parameters of this distribution. However, the posterior distribution is still defined with an integration expression which makes the closed form of the posterior distribution analytically intractable. Thus we should appeal to approximate methods. The two most prominent strategies in statistics and machine learning are Markov chain Monte Carlo (MCMC) sampling and *variational inference* (VI). In MCMC sampling, we construct a Markov chain over the hidden variables whose stationary distribution is the posterior of interest (e.g., [19], [20, Ch. 11]). Thus, we can collect samples from the exact posterior while the approximation arises from the use of finite number of samples due to the limited amount of computational resource [20, Ch. 11]. However, MCMC methods can be slow to converge and their convergence can be difficult to diagnose. This practically often limits their use to small-scale problems. Variational inference provides an alternative to computationally costly sampling-based approaches. In VI, a flexible family of distributions is defined over the hidden variables, indexed by free parameters [21]. Then a particular setting of the parameters (i.e., the member of the family) is found that is closest to the posterior, where the closeness is measured using Kullback–Leibler divergence. Thus the inference problem turns into an optimization problem [20, Ch. 10].

This paper contributes to the Bayesian study of the Watson distribution (both real and complex) and its mixture modeling. We proceed by describing the model and explaining the difficulties associated with computing the posterior distributions. We then present an analytically tractable alternative for parameter estimation based on VI. The proposed approach retains distributional information over all parameters and requires no iterative numerical calculation during the optimization procedure. Furthermore, the approach provides a lower bound on the marginal likelihood. We compare the approximate posterior distribution by VI with the exact posterior obtained by a computationally costly sampling-based approach and discuss the amount of systematic bias through toy experiments on synthetic data. Moreover, the issue of model pruning of the mixture components will be discussed.

In the experiments, we consider two rather different application domains: clustering of gene microarray data sets and blind source separation of convolutive mixtures of speech signals. Specifically, we show how formulating the mixture modeling of complex Watson distributions with VI can contribute in better performance of conventional methods, partially due to the capability of VI approach in regulating its model complexity.

The paper is organized as follows. We briefly review the Watson distribution in Section 2. In Section 3, we derive the Bayesian inference for mixture modeling with Watson distributions with variational inference. Section 4 includes empirical evaluations and discussions. The experimental results on real data are presented in Sections 5 and 6. Finally, Section 7 concludes the paper.

2 The Watson Distribution

Let $\mathbf{x} = [x_1, \dots, x_d]^{\mathrm{T}}$ be a point on the $(d-1)$-dimensional unit-hypersphere $\mathbb{S}^{d-1}$, that is $\pm\mathbf{x} \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} = \{\mathbf{x} \,|\, \mathbf{x} \in \{\mathbb{C}^d, \mathbb{R}^d\}, \|\mathbf{x}\|_2 = 1\}$. The Watson distribution is one of the commonly used distributions for modeling such data. Note that indeed any two points on the unit sphere are different but the Watson distribution owns the interesting characteristics of rotational invariance.

The distribution has density

$$\mathcal{W}(\mathbf{x} \,|\, \boldsymbol{\mu}, \lambda) = \mathrm{c}_p(\lambda) \exp\left(\lambda |\boldsymbol{\mu}^{\mathrm{H}} \mathbf{x}|^2\right), \tag{1}$$

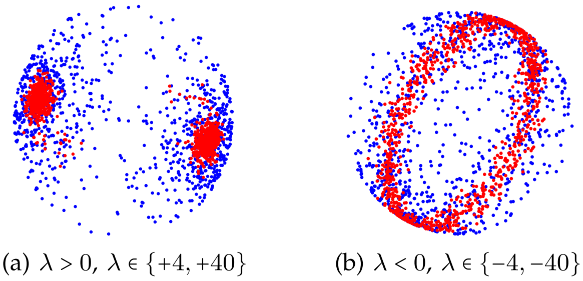

(a) $\lambda > 0$, $\lambda \in \{+4, +40\}$ (b) $\lambda < 0$, $\lambda \in \{-4, -40\}$

Fig. 1. Scatter plot of samples from four different Watson distributions on the sphere for positive and negative concentration parameters, λ. Samples generated from Watson distribution with $\lambda = \pm 4$ are shown with blue dots and those with $\lambda = \pm 40$ are shown with red dots.

where the normalization factor $\mathrm{c}_p(\lambda)$ is expressed in terms of Kummer's confluent hypergeometric function ${}_1F_1$ (e.g., [22, formula (2.1.2)], or [23, Ch. 13]) as

$$\mathrm{c}_p(\lambda) = \frac{\Gamma(p)}{(2\pi)^p {}_1F_1(r,p,\lambda)}, \begin{cases} \mathbf{x} \in \mathbb{R}, & r = 1/2, p = d/2 \\ \mathbf{x} \in \mathbb{C}, & r = 1, p = d \end{cases} \tag{2}$$

where ${}_1F_1(r,p,\lambda) = \sum_{j \geq 0} \frac{r^{(j)}}{p^{(j)}} \frac{\lambda^j}{j!}$ and $r^{(j)} \equiv \frac{\Gamma(r+j)}{\Gamma(r)}$ is the rising factorial. The distribution is parameterized by a *mean direction* $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ and a *concentration parameter* $\lambda \in \mathbb{R}$ (e.g., [1, Ch. 9.4], [7]). The real Watson distribution is useful in situations where girdle and bipolar distributions on the sphere are of interest [7]. For example in data mining, spectral embeddings show bipolar characteristics which have been suggested to be well modeled by real Watson distributions [9]. The complex Watson distribution has well established applications in statistical shape analysis [7].

When $\lambda \to 0$, $\mathcal{W}(\mathbf{x} \,|\, \boldsymbol{\mu}, \lambda)$ reduces to the uniform density, and as $\lambda \to \infty$, $\mathcal{W}(\mathbf{x} \,|\, \boldsymbol{\mu}, \lambda)$ tends to a point density (Fig. 1a). When $\lambda \to -\infty$, the density concentrates around the great circle orthogonal to the mean direction [1, Ch. 9.4] (Fig. 1b). Fig. 1 shows a scatter plot of samples drawn from a Watson distribution on the sphere.

In this work, we only focus on the positive values of concentration parameters $\lambda > 0$ as it is often the typical case in practical applications.

2.1 Prior Distribution

Let $\underline{\mathbf{X}} = \{\mathbf{x}_n\}_{n=1}^N$ denote a set of independent and identically distributed (i.i.d) observations from $\mathcal{W}(\mathbf{x}_n \,|\, \boldsymbol{\mu}, \lambda)$ for $\lambda > 0$. The Watson distribution belongs to the exponential family, thus we can take the corresponding standard conjugate prior for this family. The joint distribution $p(\boldsymbol{\mu}, \lambda, \underline{\mathbf{X}})$ is proportional to the un-normalized likelihood of $\boldsymbol{\mu}$ and λ, $h(\boldsymbol{\mu}, \lambda)$, that is $p(\boldsymbol{\mu}, \lambda, \underline{\mathbf{X}}) \propto h(\boldsymbol{\mu}, \lambda) = \exp(N \log \mathrm{c}_p(\lambda) + \lambda \boldsymbol{\mu}^{\mathrm{H}} \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^{\mathrm{H}} \boldsymbol{\mu})$. Thus, the conjugate prior for $\boldsymbol{\mu}$ and λ can be expressed as

$$p(\boldsymbol{\mu}, \lambda) \propto \exp\big(c \log \mathrm{c}_p(\lambda) + \beta_0 \lambda |\mathbf{m}_0^{\mathrm{H}} \boldsymbol{\mu}|^2\big), \tag{3}$$

where $c, \beta_o > 0$, and $(\cdot)^{\mathrm{H}}$ shows the conjugate transpose operation. We opt to factorize $p(\boldsymbol{\mu}, \lambda)$ as

$$p(\boldsymbol{\mu}, \lambda) = p(\boldsymbol{\mu} \,|\, \lambda) p(\lambda). \tag{4}$$

Considering the form of the conjugate prior in (3) and our choice of factorization in (4), the conditional distribution of $\boldsymbol{\mu}$ given λ can be expressed as a Watson distribution in the form of $\mathcal{W}(\boldsymbol{\mu} \,|\, \mathbf{m}_0, \beta_0 \lambda)$. As the result of this choice, the conjugate prior for λ, $p(\lambda)$, exists up to an analytically intractable normalization factor. In Section 3.1, we explicitly assume a Gamma distribution over $p(\lambda)$ as the prior distribution. As we shall see later in Section 3.4, the posterior distribution of λ is forced to be a Gamma distribution. Since we only address the case of positive concentration parameters, $\lambda > 0$, the choice of Gamma distribution seems intuitively reasonable; furthermore, Gamma distributions are relatively flexible and simple to work with.

2.2 Posterior Distribution

Given observed data $\underline{\mathbf{X}}$ and the assigned conjugate prior (3), the posterior distribution of $\boldsymbol{\mu}$ and λ takes the form

$$\begin{aligned} p(\boldsymbol{\mu}, \lambda \,|\, \underline{\mathbf{X}}) &= \frac{h(\boldsymbol{\mu}, \lambda)}{\int h(\boldsymbol{\mu}, \lambda) \mathrm{d}\boldsymbol{\mu} \mathrm{d}\lambda} \\ &\propto \exp\big((c+N) \log \mathrm{c}_p(\lambda) + \beta \lambda |\mathbf{m}^{\mathrm{H}} \boldsymbol{\mu}|^2\big), \end{aligned} \tag{5}$$

where β and $\mathbf{m}$ are obtained by combining the conjugate prior and the observed data. The normalizing constant in (5) is not available in a closed form, thus while this conjugate prior exists, moments and other characteristics of the corresponding posterior have to be computed numerically, e.g., by sampling from the posterior distribution. Generally speaking, sampling methods can be computationally demanding which hinder some of their practical use in large scale problems [20, Ch. 10]. We proceed by presenting an analytically tractable alternative that can approximate the posterior in a closed form. Next, through synthetic experiments, we empirically evaluate the proposed approach.

3 Bayesian Inference with VI

Variational inference is an alternative to computationally costly sampling-based approaches, which replaces sampling with optimization and provides a deterministic approximation to the posterior distribution [21], [20, Ch. 10]. This section addresses the parameter estimation of the Watson mixture model with VI.

3.1 Model Description

As before, let $\underline{\mathbf{X}} = \{\mathbf{x}_n\}_{n=1}^N$ denote a set of N i.i.d observations on the unit hypersphere. With K mixture components, the probability density function of the Watson mixture model is represented as

$$p(\underline{\mathbf{X}} \,|\, \underline{\boldsymbol{\mu}}, \underline{\lambda}, \underline{\tau}) = \prod_{n=1}^N \sum_{k=1}^K \tau_k \mathcal{W}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_k, \lambda_k), \tag{6}$$

where $\underline{\tau} = \{\tau_k\}_{k=1}^K$ ($\tau_k > 0$, $\sum_{k=1}^K \tau_k = 1$) denotes a set of mixture weights, $\underline{\boldsymbol{\mu}} = \{\boldsymbol{\mu}_k\}_{k=1}^K$ denotes a set of mean directions, and $\underline{\lambda} = \{\lambda_k\}_{k=1}^K$ denotes a set of concentration parameters. We follow the graphical model presented in Fig. 2. As shown in the model, associated with each observation $\mathbf{x}_n$, there is a corresponding latent variable $\mathbf{z}_n = [z_{n1}, \ldots, z_{nK}]^{\mathrm{T}}$ consisting of 1-of-K binary vector with elements z_{nk}, with a single element $z_{nk} = 1$ indicating that $\mathbf{x}_n$ was generated by

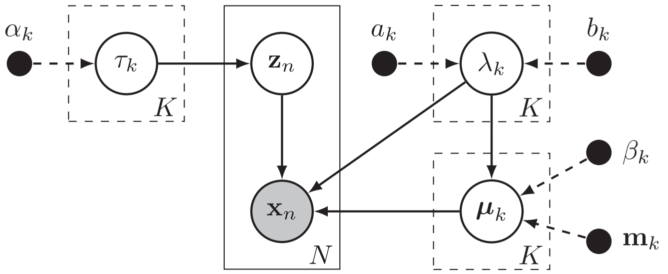


Fig. 2. Directed acyclic graph representing the Bayesian mixture model of Watson distributions. Nodes denote random variables, edges denote possible dependence, and plates denote replication.

the kth component. Indicating the set of latent variables by $\underline{\mathbf{Z}} = \{\mathbf{z}_n\}_{n=1}^N$, the conditional distribution of $\underline{\mathbf{Z}}$ given $\underline{\tau}$ is

$$p(\underline{\mathbf{Z}} \,|\, \underline{\tau}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \tau_k^{z_{nk}}. \tag{7}$$

The conditional distribution of $\underline{\mathbf{X}}$ given $\underline{\mathbf{Z}}, \underline{\boldsymbol{\mu}}, \underline{\lambda}$ is

$$p(\underline{\mathbf{X}} \,|\, \underline{\mathbf{Z}}, \underline{\boldsymbol{\mu}}, \underline{\lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{W}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_k, \lambda_k)^{z_{nk}}. \tag{8}$$

As $\underline{\mathbf{X}}$ is conditionally independent of $\underline{\tau}$ given $\underline{\mathbf{Z}}$, the conditional distribution of $\underline{\mathbf{X}}$ and $\underline{\mathbf{Z}}$ given $\underline{\boldsymbol{\mu}}, \underline{\lambda}$ can be written as

$$p(\underline{\mathbf{X}}, \underline{\mathbf{Z}} \,|\, \underline{\boldsymbol{\mu}}, \underline{\lambda}, \underline{\tau}) \quad = \textstyle\prod_{n=1}^{N} \prod_{k=1}^{K} (\tau_k \; \mathcal{W}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_k, \lambda_k))^{z_{nk}}. \tag{9}$$

Note that the mixture model in (6) can be obtained by marginalizing out the latent variables from (9) using (7).

Next, we need to introduce prior distributions over the model parameters $\underline{\boldsymbol{\mu}}, \underline{\lambda}, \underline{\tau}$. We assign a Dirichlet prior distribution for the mixture weights with the probability density function

$$\begin{aligned} p(\underline{\tau}) &= \mathrm{Dir}(\underline{\tau}|\boldsymbol{\alpha}_0) \\ &= C(\boldsymbol{\alpha}_0)\prod_{k=1}^{K} \tau_k^{\alpha_{0,k}-1}, \;\; C(\boldsymbol{\alpha}_0) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_{0,k}\right)}{\prod_{k=1}^{K}\Gamma(\alpha_{0,k})}, \end{aligned} \tag{10}$$

where $\boldsymbol{\alpha}_0 = (\alpha_{0,1}, \ldots, \alpha_{0,K})$. As we shall see later in Section 3.4, the Dirichlet prior for $\underline{\tau}$ is a conjugate prior. Motivated by our discussion in Section 2.1, we assign a Watson-Gamma prior for the mean directions and the concentration parameters as

$$p(\underline{\boldsymbol{\mu}}, \underline{\lambda}) = \prod_{k=1}^{K} \mathcal{W}(\boldsymbol{\mu}_k \,|\, \mathbf{m}_{0,k}, \beta_{0,k}\lambda_k) \;\, \mathcal{G}(\lambda_k \,|\, a_{0,k}, b_{0,k}). \tag{11}$$

In the above, $\mathcal{G}(\lambda_k|a_{0,k}, b_{0,k})$ is a Gamma density function with the shape parameter $a_{0,k}$ and the inverse scale parameter $b_{0,k}$, given by $\mathcal{G}(\lambda_k|a_{0,k}, b_{0,k}) = \frac{1}{\Gamma(a_{0,k})} b_{0,k}^{a_{0,k}} \lambda_k^{a_{0,k}-1} e^{-b_{0,k}\lambda_k}$, where $\Gamma(\cdot)$ denotes the Gamma function.

3.2 Variational Inference

In order to formulate the variational treatment of the model, we start with writing down the joint distribution of all random variables in the model. This is given by

$$p(\underline{\mathbf{X}}, \underline{\mathbf{Z}}, \underline{\tau}, \underline{\boldsymbol{\mu}}, \underline{\lambda}) = p(\underline{\mathbf{X}} \,|\, \underline{\mathbf{Z}}, \underline{\boldsymbol{\mu}}, \underline{\lambda}) p(\underline{\boldsymbol{\mu}}, \underline{\lambda}) p(\underline{\mathbf{Z}} \,|\, \underline{\tau}) p(\underline{\tau}), \tag{12}$$

in which the various factors are previously defined in Section 3.1. For convenience, we denote a set of all latent variables and parameters by $\underline{S} = \{\underline{\tau}, \underline{\boldsymbol{\mu}}, \underline{\lambda}, \underline{\mathbf{Z}}\}$. The posterior distribution of $\underline{S}$ is

$$p(\underline{S} \,|\, \underline{\mathbf{X}}) = \exp(\log\; p(\underline{\mathbf{X}}, \underline{S}) - \log\; p(\underline{\mathbf{X}})). \tag{13}$$

Working directly with this posterior is often precluded by the the need to compute the log marginal likelihood (LML) of the observations, $\log\; p(\underline{\mathbf{X}})$ (i.e., the normalizing constant). The idea behind variational free-energy methods is to approximate $p(\underline{S} \,|\, \underline{\mathbf{X}})$ with a distribution $q(\underline{S})$ that belongs to a constrained family of distributions, indexed by a variational parameter [21]. The goal is to choose a member of that family which is as close as possible to the exact posterior distribution. In VI, this closeness is measured by Kullback-Leibler (KL) divergence between $p(\underline{S} \,|\, \underline{\mathbf{X}})$ and $q(\underline{S})$. As the minimization of KL divergence is difficult, due to need of knowing the distribution being approximated, in VI a lower bound is defined on the log marginal likelihood (evidence) as $\log\; p(\underline{\mathbf{X}}) = \mathcal{L}[q] + \mathrm{KL}[q(\underline{S}) \| p(\underline{S} \,|\, \underline{\mathbf{X}})]$, where

$$\mathcal{L}[q] = \mathbb{E}_q[\log\; p(\underline{\mathbf{X}}, \underline{S})] - \mathbb{E}_q[\log\; q(\underline{S})] \tag{14}$$

is known as the evidence lower bound. Operator $\mathbb{E}_q[\cdot]$ takes the expectation of variables in its argument with respect to the variational variable distribution $q(\cdot)$. Since $p(\underline{\mathbf{X}})$ is constant with respect to the family q, the lower bound is maximized when $p(\underline{S} \,|\, \underline{\mathbf{X}}) = q(\underline{S})$. Typically, q will be constrained to a family of simpler distributions, by breaking some of dependencies between variables that make the true posterior difficult to compute. The choice of factorization techniques affects the quality of the approximation (e.g., this issue has been discussed in [24], [25]). The simplest variational family of distributions is the *mean-field* family (e.g., [26]). In this family, each hidden variable is independent and governed by its own parameter. This means the variational posterior $q(\underline{S})$ can be explicitly expressed by $q(\underline{S}) = q(\underline{\mathbf{Z}}) q(\underline{\tau}) q(\underline{\boldsymbol{\mu}}) q(\underline{\lambda})$. However, we consider a particular form of factorization known as *structural factorization*[1] of the variational posterior as

$$q(\underline{S}) = q(\underline{\mathbf{Z}}) q(\underline{\tau}) q(\underline{\boldsymbol{\mu}} \,|\, \underline{\lambda}) q(\underline{\lambda}), \tag{15}$$

by which we can retain some information about uncertainty and allow this information to propagate between factors $\underline{\boldsymbol{\mu}}$ and $\underline{\lambda}$. Thus, unlike the mean-field approximation which violates any possible dependencies between $\underline{\boldsymbol{\mu}}$ and $\underline{\lambda}$, the structural factorization can partly preserve the actual dependence.

The corresponding sequential update equations for these factors can be derived by maximizing the lower bound on the log marginal likelihood. We start with rewriting $\mathcal{L}(q)$, (14), as

$$\mathcal{L}(q) = \mathbb{E}_q\left[\log \frac{p(\underline{\mathbf{X}} \,|\, \underline{\mathbf{Z}}, \underline{\boldsymbol{\mu}}, \underline{\lambda}) p(\underline{\boldsymbol{\mu}} \,|\, \underline{\lambda}) p(\underline{\lambda}) p(\underline{\mathbf{Z}} \,|\, \underline{\tau}) p(\underline{\tau})}{q(\underline{\mathbf{Z}}) q(\underline{\boldsymbol{\mu}} \,|\, \underline{\lambda}) q(\underline{\lambda}) q(\underline{\tau})}\right]. \tag{16}$$

Evaluation of this bound requires intractable moment computations. For example, for evaluation of $\mathbb{E}_q\big[\log\; p(\underline{\mathbf{X}} \,|\, \underline{\mathbf{Z}}, \underline{\boldsymbol{\mu}}, \underline{\lambda})\big]$,

[1] The structural factorization has been also considered in derivation of variational inference for Gaussian mixture models [20, Ch. 10.2].

one needs to compute $\mathbb{E}_q[\log\ {}_1F_1(r,p,\lambda_k)]$ and $\mathbb{E}_q[\lambda_k|\boldsymbol{\mu}_k^{\mathrm{H}}\mathbf{x}_n|^2]$ which are not available in closed forms. Consequently, as we shall see later, optimization of the variational distribution $q(\underline{\mathbf{Z}})$ and $q(\underline{\lambda})$ can not be carried out with standard computations in the exponential family. One remedy to avoid this intractable moment computation is to further bound the lower bound $\mathcal{L}(q)$ through a Taylor series expansion. This bound is tight at one point in the parameter distribution. Similar ideas have been previously studied in literature (e.g., [27], [28], [29]).

We proceed by defining proper "help functions" and deriving necessary bounds for their moment computations.

3.3 Help Functions

To derive an approximate solution which does not violate the lower bound, we need to define the following help functions. The convexity (or concavity) of the introduced help functions are examined through the following Lemmas. Proofs are provided as an Appendix in the supplementary materials, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2015.2498935.

Lemma 3.1. *The function* ${}_1F_1(\nu,\kappa,x)$ *is log-convex relative to* $\log\ x$ *for* $\kappa > \nu > 0$ *and* $x > 0, x \in \mathbb{R}$.

Lemma 3.2. *Defining* $\psi(x) = \frac{\partial}{\partial x}\log\ {}_1F_1(\nu,\kappa,x)$, *the function* $x\psi(x)$ *is convex relative to* $\log\ x$ *for all* $\kappa > \nu > 0$ *and* $x > 0, x \in \mathbb{R}$.

Lemma 3.3. *Defining the help function*

$$ {}_1\mathcal{H}_1(\nu,\kappa,x) = x^{\kappa}\,{}_1F_1(\nu,\kappa,x), \tag{17}$$

the function ${}_1\mathcal{H}_1(\nu,\kappa,x)$ *is log-concave with respect to* x *for all* $\kappa > \nu > 0$ *and* $x > 0, x \in \mathbb{R}$.

From Lemma 3.1-3.3, we derive the following bounds:

$$\log\ {}_1\mathcal{H}_1(r,p,\lambda_k) \le \log\ {}_1\mathcal{H}_1(r,p,\bar{\lambda}_k) + \varphi(\bar{\lambda}_k)(\lambda_k - \bar{\lambda}_k), \tag{18}$$

$$\begin{aligned}\log\ {}_1\mathcal{H}_1(r,p,\beta_{0,k}\lambda_k) \le{}& \log\ {}_1\mathcal{H}_1(r,p,\beta_{0,k}\bar{\lambda}_k)\\ &+ \varphi(\beta_{0,k}\bar{\lambda}_k)(\beta_{0,k}\lambda_k - \beta_{0,k}\bar{\lambda}_k),\end{aligned} \tag{19}$$

$$\begin{aligned}\log\ {}_1F_1(r,p,\beta_k\lambda_k) \ge{}& \log\ {}_1F_1(r,p,\beta_k\bar{\lambda}_k)\\ &+ \beta_k\bar{\lambda}_k\psi(\beta_k\bar{\lambda}_k)(\log\ \beta_k\lambda_k - \log\ \beta_k\bar{\lambda}_k),\end{aligned} \tag{20}$$

$$\begin{aligned}\lambda_k\psi(\beta_k\lambda_k) \ge{}& \bar{\lambda}_k\psi(\beta_k\bar{\lambda}_k) + \bar{\lambda}_k\big(\psi(\beta_k\bar{\lambda}_k)\\ &+ \beta_k\bar{\lambda}_k\psi'(\beta_k\bar{\lambda}_k)\big)(\log\ \beta_k\lambda_k - \log\ \beta_k\bar{\lambda}_k),\end{aligned} \tag{21}$$

where we have defined $\psi'(x) = \frac{\partial}{\partial x}\psi(x)$ and $\varphi(x) = \frac{\partial}{\partial x}\log\ {}_1\mathcal{H}_1(\nu,\kappa,x)$. In (18)-(21), the bounds are tight at $\lambda_k = \bar{\lambda}_k$.

3.4 Variational Posterior Distribution

In this section we optimize the variational posterior distributions (15) by maximizing the lower bound (16) with respect to the variational parameters. An explicit algorithm will be presented in this regard.

3.4.1 Variational Expectation

We rewrite (16) by including all terms which do not involve $\underline{\mathbf{Z}}$ into a constant term as

$$\begin{aligned}\mathcal{L}(q) &= \mathrm{cst} + \mathbb{E}_q\left[\log\ \frac{p(\underline{\mathbf{X}}\,|\,\underline{\mathbf{Z}},\underline{\boldsymbol{\mu}},\underline{\lambda})p(\underline{\mathbf{Z}}\,|\,\underline{\tau})}{q(\underline{\mathbf{Z}})}\right]\\ &= \mathrm{cst} + \sum_{k=1}^{K}\sum_{n=1}^{N}\log\ \rho_{nk}\mathbb{E}_q[z_{nk}] - \mathbb{E}_q[\log\ q(\underline{\mathbf{Z}})],\end{aligned} \tag{22}$$

where $\log\ \rho_{nk}$ is given by

$$\begin{aligned}\log\ \rho_{nk} ={}& -\mathbb{E}_q[\log\ {}_1\mathcal{H}_1(r,p,\lambda_k)] + \mathbb{E}_q\Big[\lambda_k|\boldsymbol{\mu}_k^{\mathrm{H}}\mathbf{x}_n|^2\Big]\\ &+ \mathbb{E}_q[\log\ \tau_k] - p\log\ 2\pi + \log\ \Gamma(p) + p\mathbb{E}_q[\log\ \lambda_k],\end{aligned} \tag{23}$$

where we have used (17). We observe that $\log\ \rho_{nk}$ is not tractable as it involves intractable moments $\mathbb{E}_q[\log\ {}_1\mathcal{H}_1(r,p,\lambda_k)]$ and $\mathbb{E}_q\big[\lambda_k|\boldsymbol{\mu}_k^{\mathrm{H}}\mathbf{x}_n|^2\big]$. From (18), it is easy to show

$$\begin{aligned}\mathbb{E}_q[\log\ {}_1\mathcal{H}_1(r,p,\lambda_k)] \le{}& \log\ {}_1\mathcal{H}_1(r,p,\bar{\lambda}_k)\\ &+ \varphi(\bar{\lambda}_k)(\mathbb{E}_q[\lambda_k] - \bar{\lambda}_k).\end{aligned} \tag{24}$$

To evaluate $\mathbb{E}_q[\lambda_k|\boldsymbol{\mu}_k^{\mathrm{H}}\mathbf{x}_n|^2]$, first we need to evaluate $\mathbb{E}_q\big[\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{H}}\big]$ which is equivalent to evaluating the second moment of the Watson distribution, as we expect $q(\boldsymbol{\mu}_k\,|\,\lambda_k)$ to have a Watson distribution. Since the Watson distribution is a special case of the Bingham distribution [1, Ch. 9], we adopt the approach proposed in [30] for computation of the second moment of the Bingham distribution and employ it, with slight adaptation, for computation of the second moment of the Watson distribution. Thus, we get

$$\begin{aligned}&\mathbb{E}_q\big[\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{H}}\big] = \vartheta(\beta_k\lambda_k)\mathbf{m}_k\mathbf{m}_k^{\mathrm{H}},\\ &\vartheta(\beta_k\lambda_k) = \frac{\partial}{\partial\beta_k\lambda_k}\left(\log\ \frac{1}{\mathrm{c}_p(\beta_k\lambda_k)}\right) = \psi(\beta_k\lambda_k).\end{aligned} \tag{25}$$

Taking into account (21) and (25), $\mathbb{E}_q[\lambda_k|\boldsymbol{\mu}_k^{\mathrm{H}}\mathbf{x}_n|^2]$ is bounded as

$$\begin{aligned}&\mathbb{E}_q\big[\lambda_k|\boldsymbol{\mu}_k^{\mathrm{H}}\mathbf{x}_n|^2\big] = \mathbb{E}_q[\lambda_k\psi(\beta_k\lambda_k)]|\mathbf{m}_k^{\mathrm{H}}\mathbf{x}_n|^2\\ &\ge \big(\bar{\lambda}_k\psi(\beta_k\bar{\lambda}_k) + \bar{\lambda}_k\big(\psi(\beta_k\bar{\lambda}_k) + \beta_k\bar{\lambda}_k\psi'(\beta_k\bar{\lambda}_k)\big)\\ &\big(\mathbb{E}_q[\log\ \lambda_k] + \log\ \beta_k - \log\ \beta_k\bar{\lambda}_k\big)\big)|\mathbf{m}_k^{\mathrm{H}}\mathbf{x}_n|^2.\end{aligned} \tag{26}$$

Finally by considering (24) and (26) in (23), $\log\ \rho_{nk}$ is bounded as

$$\begin{aligned}\log\ \rho_{nk} \ge{}& \widetilde{\log\ \rho_{nk}}\\ \triangleq{}& \mathbb{E}_q[\log\ \tau_k] - p\log\ 2\pi + \log\ \Gamma(p) + p\mathbb{E}_q[\log\ \lambda_k]\\ &- \log\ {}_1\mathcal{H}_1(r,p,\bar{\lambda}_k) - \varphi(\bar{\lambda}_k)(\mathbb{E}_q[\lambda_k] - \bar{\lambda}_k)\\ &+ \bar{\lambda}_k\psi(\beta_k\bar{\lambda}_k) + \big(\bar{\lambda}_k\big(\psi(\beta_k\bar{\lambda}_k) + \beta_k\bar{\lambda}_k\psi'(\beta_k\bar{\lambda}_k)\big)\\ &\big(\mathbb{E}_q[\log\ \lambda_k] + \log\ \beta_k - \log\ \beta_k\bar{\lambda}_k\big)\big)|\mathbf{m}_k^{\mathrm{H}}\mathbf{x}_n|^2.\end{aligned} \tag{27}$$

Thus, we can further lower bound the exact lower bound $\mathcal{L}(q)$ as,

$$\begin{aligned}\mathcal{L}(q) \ge \dot{\mathcal{L}}(q) \triangleq{}& \mathrm{cst}\\ &+ \sum_{k=1}^{K}\sum_{n=1}^{N}\big(\mathbb{E}_q\big[z_{nk}\widetilde{\log\ \rho_{nk}}\big]\big) - \mathbb{E}_q[\log\ q(\underline{\mathbf{Z}})],\end{aligned} \tag{28}$$

where the bound is tight at one point in the parameter distribution (that is $\lambda_k = \bar{\lambda}_k$). Inequality (28) implies that our approximation by (27) does not violate the lower bound. Thus, the variational posterior distribution of latent variables can be expressed as

$$\log\ q^*(\underline{\mathbf{Z}}) = \mathrm{cst} + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\widetilde{\log\ \rho_{nk}}. \tag{29}$$

Hence, we get $q^*(\underline{\mathbf{Z}}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\xi_{nk}^{z_{nk}}$, which is a categorical distribution with

$$\xi_{nk} = \frac{\mathrm{e}^{\left(\widetilde{\log\ \rho_{nk}}\right)}}{\sum_{j=1}^{K}\mathrm{e}^{\left(\widetilde{\log\ \rho_{nj}}\right)}}, \tag{30}$$

where ξ_{nk} are nonnegative and have a unit sum. The quantities ξ_{nk} play the role of *responsibilities*. For $q^*(\underline{\mathbf{Z}})$, we have $\mathbb{E}_q[z_{nk}] = \xi_{nk}$.

3.4.2 Variational Maximization

Starting with optimization of the posterior distribution of the mixture weights $q(\underline{\tau})$, we rewrite (16) by including all terms which do not involve $\underline{\tau}$ into a constant term

$$\begin{aligned}\mathcal{L}(q) &= \mathrm{cst} + \mathbb{E}_q\left[\log\ \frac{p(\underline{\mathbf{Z}}\,|\,\underline{\tau})p(\underline{\tau})}{q(\underline{\tau})}\right]\\ &= \mathrm{cst} + \sum_{k=1}^{K}\mathbb{E}_q[f(\tau_k)] - \mathbb{E}_q[\log\ q(\underline{\tau})],\end{aligned} \tag{31}$$

where $f(\tau_k) = ((\alpha_{0,k}-1) + \sum_{n=1}^{N}\xi_{nk})\log\ \tau_k$. Hence, we have $\log\ q^*(\underline{\tau}) = \mathrm{cst} + \sum_{k=1}^{K} f(\tau_k)$. Taking exponential of both sides of this expression, $q^*(\underline{\tau})$ is recognized to have a Dirichlet distribution,

$$q^*(\underline{\tau}) = \mathrm{Dir}(\underline{\tau}|\boldsymbol{\alpha}),\quad \alpha_k = \alpha_{0,k} + \sum_{n=1}^{N}\xi_{nk}. \tag{32}$$

Next, we consider optimization of $q(\underline{\boldsymbol{\mu}}, \underline{\lambda})$ which comprises a sum over k of terms involving $\boldsymbol{\mu}_k$ and λ_k leading to the further factorization $q(\underline{\boldsymbol{\mu}}, \underline{\lambda}) = \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \lambda_k)$. The variational posterior distribution $q(\boldsymbol{\mu}_k, \lambda_k)$ can be written in the form of $q(\boldsymbol{\mu}_k, \lambda_k) = q(\boldsymbol{\mu}_k\,|\,\lambda_k)q(\lambda_k)$. Considering terms in (16) involving only $\underline{\boldsymbol{\mu}}$ and $\underline{\lambda}$,

$$\begin{aligned}\mathcal{L}(q) &= \mathrm{cst} + \mathbb{E}_q\left[\log\ \frac{p(\underline{\mathbf{X}}\,|\,\underline{\mathbf{Z}}, \underline{\boldsymbol{\mu}}, \underline{\lambda})p(\underline{\boldsymbol{\mu}}\,|\,\underline{\lambda})p(\underline{\lambda})}{q(\underline{\boldsymbol{\mu}}\,|\,\underline{\lambda})q(\underline{\lambda})}\right]\\ &= \mathrm{cst} + \sum_{k=1}^{K}\mathbb{E}_q[f(\boldsymbol{\mu}_k, \lambda_k)] - \mathbb{E}_q[\log\ q(\boldsymbol{\mu}_k, \lambda_k)],\end{aligned} \tag{33}$$

$$\begin{aligned}f(\boldsymbol{\mu}_k, \lambda_k) &= (a_{0,k}-1)\log\ \lambda_k - b_{0,k}\lambda_k\\ &\quad + \sum_{n=1}^{N}\xi_{nk}\Big(-\log\ {}_1F_1(r, p, \lambda_k) + \lambda_k|\boldsymbol{\mu}_k^{\mathrm{H}}\mathbf{x}_n|^2\Big)\\ &\quad - \log\ {}_1F_1(r, p, \beta_{0,k}\lambda_k) + \beta_{0,k}\lambda_k|\mathbf{m}_{0,k}^{\mathrm{H}}\boldsymbol{\mu}_k|^2.\end{aligned} \tag{34}$$

Inspecting (34) and reading off those terms involving only $\boldsymbol{\mu}_k$, we can express $\log\ q^*(\boldsymbol{\mu}_k\,|\,\lambda_k)$ as

$$\begin{aligned}\log\ q^*(\boldsymbol{\mu}_k\,|\,\lambda_k) &= \mathrm{cst}\\ &\quad + \beta_{0,k}\lambda_k|\mathbf{m}_{0,k}^{\mathrm{H}}\boldsymbol{\mu}_k|^2 + \sum_{n=1}^{N}\xi_{nk}\lambda_k|\boldsymbol{\mu}_k^{\mathrm{H}}\mathbf{x}_n|^2.\end{aligned} \tag{35}$$

By taking exponential of both sides of (35), $q^*(\boldsymbol{\mu}_k\,|\,\lambda_k)$ is recognized to be a Watson density

$$q^*(\boldsymbol{\mu}_k\,|\,\lambda_k) = \mathcal{W}(\boldsymbol{\mu}_k\,|\,\mathbf{m}_k, \beta_k\lambda_k), \tag{36}$$

$$\beta_k = \kappa_k^*,\qquad \kappa_k^* = \kappa_k^1 \geq \cdots \geq \kappa_k^p \tag{37}$$

$$\mathbf{m}_k = \boldsymbol{\upsilon}_k^*,\qquad \boldsymbol{\upsilon}_k^* = \boldsymbol{\upsilon}_k^1 \in \{\boldsymbol{\upsilon}_k^1, \ldots, \boldsymbol{\upsilon}_k^p\}, \tag{38}$$

where κ_k^* is the largest eigenvalue of the positive definite matrix $\left(\beta_0\mathbf{m}_{0,k}\mathbf{m}_{0,k}^{\mathrm{H}} + \sum_{n=1}^{N}\xi_{nk}\mathbf{x}_n\mathbf{x}_n^{\mathrm{H}}\right)$, and $\boldsymbol{\upsilon}_k^*$ is the eigenvector corresponding to this eigenvalue.[2]

Now, we can determine $\log\ q^*(\lambda_k)$ simply as $\log\ q^*(\lambda_k) = \log\ q^*(\boldsymbol{\mu}_k, \lambda_k) - \log\ q^*(\boldsymbol{\mu}_k|\lambda_k)$, where on the right-hand side of this equality, we substitute for $\log\ q^*(\boldsymbol{\mu}_k|\lambda_k)$ using (35) and for $\log\ q^*(\boldsymbol{\mu}_k, \lambda_k)$ using (34). Retaining those terms that have some functional dependence on λ_k and absorbing other terms into a constant term, the lower bound (16) can be expressed as

$$\mathcal{L}(q) = \mathrm{cst} + \sum_{k=1}^{K}\big(\mathbb{E}_q[f(\lambda_k)]\big) - \mathbb{E}_q[\log\ q(\underline{\lambda})], \tag{39}$$

$$\begin{aligned}f(\lambda_k) &= (a_{0,k}-1)\log\ \lambda_k - b_{0,k}\lambda_k + p\sum_{n=1}^{N}\xi_{nk}\log\ \lambda_k\\ &\quad - \sum_{n=1}^{N}\xi_{nk}\log\ {}_1\mathcal{H}_1(r, p, \lambda_k) + p\ \log\ \beta_{0,k}\lambda_k\\ &\quad - \log\ {}_1\mathcal{H}_1(r, p, \beta_{0,k}\lambda_k) + \log\ {}_1F_1(r, p, \beta_k\lambda_k),\end{aligned} \tag{40}$$

where we have used (17). We note that (39) is not analytically tractable as (40) involves intractable moments. Considering (18) and (19) in (40), $f(\lambda_k)$ is bounded as

$$\begin{aligned}f(\lambda_k) \geq \widetilde{f(\lambda_k)} &\triangleq (a_{0,k}-1)\log\ \lambda_k - b_{0,k}\lambda_k\\ &\quad +\ p\sum_{n=1}^{N}\xi_{nk}\log\ \lambda_k - \sum_{n=1}^{N}\xi_{nk}\log\ {}_1\mathcal{H}_1(r, p, \bar{\lambda}_k)\\ &\quad +\ p\log\ \beta_{0,k}\lambda_k - \log\ {}_1\mathcal{H}_1(r, p, \beta_{0,k}\bar{\lambda}_k)\\ &\quad +\ \log\ {}_1F_1(r, p, \beta_k\bar{\lambda}_k) - \sum_{n=1}^{N}\xi_{nk}\varphi(\bar{\lambda}_k)(\lambda_k - \bar{\lambda}_k)\\ &\quad -\ \varphi(\beta_{0,k}\bar{\lambda}_k)(\beta_{0,k}\lambda_k - \beta_{0,k}\bar{\lambda}_k)\\ &\quad +\ \psi(\beta_k\bar{\lambda}_k)\beta_k\bar{\lambda}_k(\log\ \beta_k\lambda_k - \log\ \beta_k\bar{\lambda}_k).\end{aligned} \tag{41}$$

Thus, we can further lower bound the exact lower bound $\mathcal{L}(q)$ as,

$$\begin{aligned}\mathcal{L}(q) &\geq \ddot{\mathcal{L}}(q)\\ &\triangleq \mathrm{cst} + \sum_{k=1}^{K}\left(\mathbb{E}_q[\widetilde{f(\lambda_k)}]\right) - \mathbb{E}_q[\log\ q(\underline{\lambda})],\end{aligned} \tag{42}$$

where the bound is tight at the point $\lambda_k = \bar{\lambda}_k$ in the parameter distribution. Inequality (42) shows our approximation by (41) does not violate the lower bound. Thus,

[2] Similarly in the ML estimation, the largest eigenvector is considered as the ML estimate of the mean direction [1, Ch. 14.7].

$\log\ q^*(\lambda_k)$ can be expressed in terms of λ_k and $\log\ \lambda_k$ as

$$\log\ q^*(\lambda_k) = \mathrm{cst} + (a_k - 1)\log\ \lambda_k - b_k\lambda_k, \tag{43}$$

where we have defined

$$a_k = a_{0,k} + p\left(1 + \sum_{n=1}^{N} \xi_{nk}\right) + \beta_k\bar{\lambda}_k\psi(\beta_k\bar{\lambda}_k), \tag{44}$$

$$b_k = b_{0,k} + \sum_{n=1}^{N} \xi_{nk}\varphi(\bar{\lambda}_k) + \beta_{0,k}\varphi(\beta_{0,k}\bar{\lambda}_k). \tag{45}$$

By taking exponential of both sides of (43), we recognize $q^*(\lambda_k)$ as a Gamma density

$$q^*(\lambda_k) = \mathcal{G}(\lambda_k \,|\, a_k, b_k). \tag{46}$$

In the above expressions, $\bar{\lambda}_k = \frac{a_k}{b_k}$ that is obtained from the previous iteration. The other expectations involved in (27) are given as $\mathbb{E}_q[\log\ \lambda_k] = \Psi(a_k) - \log\ b_k$, $\mathbb{E}_q[\lambda_k] = \frac{a_k}{b_k}$, $\mathbb{E}_q[\log\ \tau_k] = \Psi(\alpha_k) - \Psi(\hat{\alpha})$, $\hat{\alpha} = \sum_{k=1}^{K} \alpha_k$, where $\Psi(y) \equiv \frac{\mathrm{d}}{\mathrm{d}y}$ $\log\ \Gamma(y)$ is known as digamma function.

3.5 Summary of the Algorithm

The algorithm starts with initialization of the variational parameters. The optimization of the variational posterior distribution $q(\underline{\mathbf{Z}})q(\underline{\tau}, \underline{\boldsymbol{\mu}}, \underline{\lambda})$ involves cycling between optimization of $q(\underline{\mathbf{Z}})$ and $q(\underline{\tau}, \underline{\boldsymbol{\mu}}, \underline{\lambda})$, which is analogous to the expectation and the maximization steps in the maximum-likelihood expectation-maximization algorithm. First, we use the current distributions over the model parameters to evaluate the responsibilities ξ_{nk}. Next, these responsibilities are employed to re-estimate the variational distribution over the parameters. The procedure is guaranteed to converge as the lower bound is convex in each of the factors [31]. A summary of the algorithm is presented in Algorithm 1.

Algorithm 1. Bayesian Estimation of Watson Mixture Model with Variational Inference

```
Input: $\underline{\mathbf{X}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ on $\mathbb{S}^{D-1}$
Step 1: Initialization
  1) Set the number of components, $K$.
  2) Set the prior distribution parameters: $\alpha_{0,k}$, $\beta_{0,k}$, $\mathbf{m}_{0,k}$, $a_{0,k}$,
     and $b_{0,k}$ (see Section 3.6).
  3) Initialize $\xi_{nk}$ by the K-means algorithm.
  4) Calculate the initial guess of $\bar{\lambda}_k$ from $\bar{\lambda}_k = \frac{a_{0,k}}{b_{0,k}}$.
repeat
    Step 2: Optimization of the Posterior Distribution
      1) Update responsibilities $\xi_{nk}$ by computing (27), (30).
      2) Update posterior distribution parameters $\alpha_k$, $\beta_k$, $\mathbf{m}_k$, $a_k$,
         $b_k$ by (32), (36), (37), (44), (45).
      3) Update $\bar{\lambda}_k$ by setting $\bar{\lambda}_k = \frac{a_k}{b_k}$.
    Step 3: Lower Bound Evaluation
    Evaluate lower bound (16). The lower bound is monitored in
    each iteration.
until convergence
```

3.6 Initialization of the Variational Parameters

In the initialization of the algorithm, when we have no prior knowledge on the data, the prior parameters $\alpha_{0,k}$, $a_{0,k}$, and $b_{0,k}$ are chosen such that the corresponding posterior distributions are influenced mostly by the data and less by the priors, i.e., $0 < \alpha_{0,k}, a_{0,k}, b_{0,k} \ll 1$ where the same prior values can be considered for all components. The prior parameter $\mathbf{m}_{0,k}$ is randomly initialized from the data such that $\|\mathbf{m}_{0,k}\|_2 = 1$, and $\beta_{0,k}$ is set to the largest eigenvalue of the positive definite matrix $\mathbf{m}_{0,k}\mathbf{m}_{0,k}^{\mathrm{H}}$ for each component. By observing a new set of data, we can initialize the prior parameters in an informative way. Although the algorithm provides a bound for different starting values of the variational parameters, poor choices of initialization can lead to local maxima that yield poor bounds. Multiple restarts can help to avoid this, with cost of slowing the overall convergence of the procedure. For increasing the convergence speed, one may consider the adaptive bound optimization method described in [32, Section 4.2].

4 Empirical Evaluations

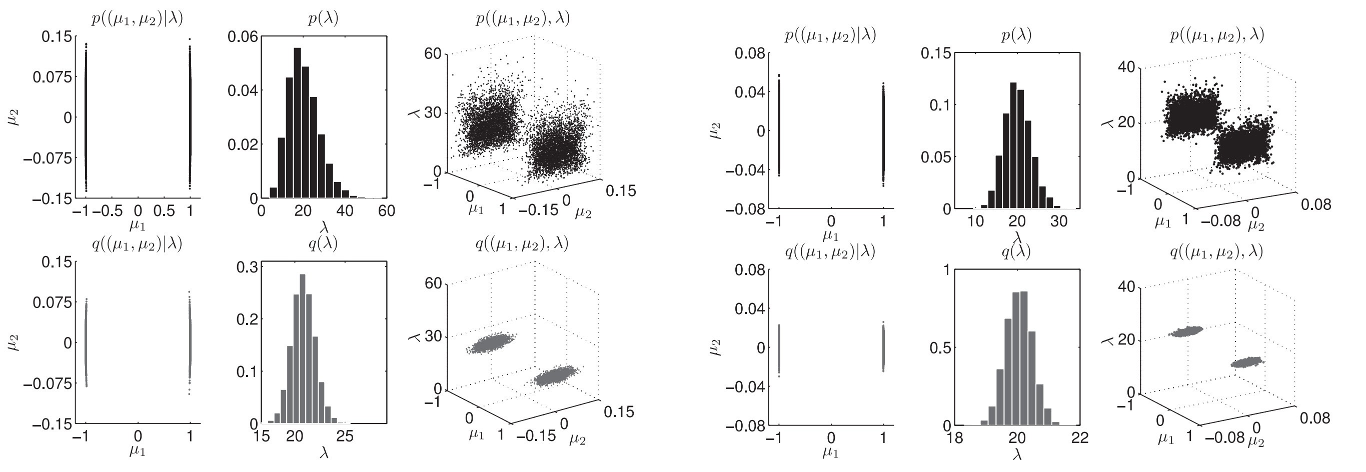
The presented variational approach has some potential advantages over MCMC approaches. It is deterministic, and offers an optimization criterion (16) which can be used to determine convergence. In contrast, evaluating convergence of an MCMC approach requires to determine when the Markov chain has reached its stationary distribution, which is a challenging task. On the other hand, there are some disadvantages along with the presented approach. First, the optimization procedure can lead to local maxima in the variational parameter space—one can mitigate the effects of local maxima using multiple restarts. Second, the variational approach yields only an approximation to the posterior and incurs an unknown bias. In contrast, MCMC approaches can potentially provide samples from the exact posterior. However, it is notable that obtaining the exact posterior highly depends on the Markov chain's convergence to its stationary distribution, which can be a slow process. In this section, we empirically examine the presented variational approach by experimental evaluations.

4.1 Posterior Distribution

It has been argued that variational inference might underestimate uncertainty which may result in compact posterior distributions [25]. However, it is notable that, as discussed by Turner et al. [33], "the compactness folk theorem is a useful rule of thumb, rather than a law of the cosmos". It has also been shown that using a lower bound to the VI criterion function, as in our (28) and (42), can lead to a bias in the parameter estimation [25]. Here, we shall show how well the variational approximation performs compared to a computationally costly sampling-based approach. For this purpose, we compare the approximate posterior distribution obtained by VI with the one obtained by a sampling approach. The posterior distribution obtained using the sampling approach is regarded as the true (exact) posterior distribution.

Here, we use the Hamiltonian Monte Carlo (HMC; [34]) in order to draw samples from the posterior distribution.[3] HMC is a widely applicable algorithm in statistics which belongs to a family of techniques for sampling in

[3] It is mentionable that the HMC only provides an approximation that is asymptotically converging to the exact posterior, with infinitely many samples.



(a) N=10

(b) N=100

Fig. 3. Comparison of the posterior distribution estimated by the HMC inference (first row) with the approximate posterior distribution estimated by the variational approach (second row) for Example 4.1: (left) the posterior distribution of the mean direction $p((\mu_1, \mu_2) \mid \lambda)$, $q((\mu_1, \mu_2) \mid \lambda)$; (middle) the posterior distribution of the concentration parameter $p(\lambda)$, $q(\lambda)$; (right) the posterior distribution of the mean direction and the concentration parameter $p((\mu_1, \mu_2), \lambda)$, $q((\mu_1, \mu_2), \lambda)$.

continuous state spaces. The method is best suited to models with a small number of parameters, and hence is chosen for our comparative study. It is important to note that the goal here is not to propose an efficient sampling approach based on HMC for the Watson model but rather to use HMC for the purpose of validation of the proposed approximate method.

The standard HMC algorithm works in two main steps, [34, Ch. 3]. In the first step, the momentum variables are randomly drawn from their independent, Gaussian proposals. In the second step, a Metropolis update is performed, using Hamiltonian dynamics to propose a new state. Hamiltonian dynamics is simulated with the leapfrog method. In this way, distant jumps can be proposed and random-walk behavior can be avoided. In fact, as discussed by Neal [34], the biggest advantage of HMC over simple random-walk Metropolis or Gibbs sampling methods is that it can propose to move to a distant point, with a high probability of acceptance. HMC performance however depends on the chosen settings of parameters in the leapfrog method, which are leapfrog stepsize ϵ and the number of leapfrog steps L. Here we use a variant of HMC with partial momentum refreshment as described in [34, Section 5.3]. A fairly good performance, in particular for the following example (Example 4.1), was achieved with long trajectories (e.g., $L = 25$) and relatively small step size ($\epsilon = 0.1$) with partial momentum refreshment of 0.9.

Example 4.1. For this illustrative example, we simulated samples of size $N \in \{10, 100\}$ from a single Watson distribution, with mean direction vector $\boldsymbol{\mu}_{\text{true}} = [0, 1]^{\mathrm{T}}$ and concentration parameter $\lambda_{\text{true}} = 20$. For generating random numbers from a given Watson distribution, we used the algorithm proposed by Hoff [35] which covers a more general case of Bingham distributions [1, Ch. 9.4]. Fig. 3 illustrates the estimated posterior distribution with HMC and VI—the one by HMC is regarded as the true posterior distribution. We observe that the approximate posterior $q((\mu_1, \mu_2), \lambda)$ by VI is concentrated in a narrower region compared to the true posterior $p((\mu_1, \mu_2), \lambda)$, which can be explained by the compactness tendency of the variational factorization. As discussed earlier, since we allow some uncertainty to propagate between factors $\boldsymbol{\mu}$ and λ, the approximate posterior has nearly preserved the actual dependence.

Example 4.2. In this example we empirically evaluate the amount of introduced systematic bias by the variational approach and the average mean square error (MSE) through a toy experiment. Note that the bias is discussed in the sense of point estimates from the posterior distribution.[4] For this purpose, we generated $I = 1{,}000$ data sets with the same size $N \in \{100, 200\}$ from the same fixed source parameters of the chosen dimensionality $d \in \{10, 20, 30, 40, 50\}$. In all experiments $\lambda_{\text{true}} = 20$ and $\boldsymbol{\mu}_{\text{true}} = [\mathbf{0}_{d-1}, 1]^{\mathrm{T}}$ where $\mathbf{0}_d$ is a zero vector of dimension $d - 1$. Next, for each data set, i, we compute the systematic bias and average MSE as follows.

Point estimates to the source parameters $\boldsymbol{\mu}$ and λ are given by the mean of the approximate posterior distribution of λ and the *central direction*[5] computed from the approximate posterior distribution of $\boldsymbol{\mu}$. These point estimates are explicitly defined for each data set i as $\hat{\lambda}_i \triangleq \mathbb{E}_q[\lambda_i]$ and $\hat{\boldsymbol{\mu}}_i$, where $\hat{\boldsymbol{\mu}}_i$ is the normalized vector in the direction of $\mathbb{E}_q[\boldsymbol{\mu}_i]$. Further let $\bar{\lambda}^{(\mathrm{VI})}$ and $\bar{\boldsymbol{\mu}}^{(\mathrm{VI})}$ denote the average of the posterior point estimates over all data sets, defined as: $\bar{\lambda}_{\mathrm{VI}} = \frac{1}{I}\sum_{i=1}^{I} \hat{\lambda}_i$ and $\bar{\boldsymbol{\mu}}^{(\mathrm{VI})} = \boldsymbol{\nu}^*$, where $\boldsymbol{\nu}^*$ is the principal eigenvector of the average matrix $\frac{1}{I}\sum_{i=1}^{I} \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{H}}$.[6] Then, the systematic bias between $\bar{\lambda}^{(\mathrm{VI})}$ and $\lambda^{(\mathrm{true})}$ can be roughly obtained by measuring the Euclidean distance between them (shown by $\operatorname{dist}(\bar{\lambda}^{(\mathrm{VI})}, \lambda^{(\mathrm{true})})$; and the systematic bias between $\bar{\boldsymbol{\mu}}^{(\mathrm{VI})}$ and

[4] One may argue that the whole point of having a Bayesian framework is to avoid taking point estimates from the posterior. However, in some applications, we may still be interested in knowing these values.

[5] Note that central direction is not equal to the mean value of $\mathbf{x}$. It is in fact a normalized vector in the same direction as the expected value of $\mathbf{x}$.

[6] Note that $\boldsymbol{\nu}^*$ is the vector that minimizes the sum of square deviations $\sum_{i=1}^{I} \sin^2\theta_i$ across all data sets, where θ_i is the angle between $\boldsymbol{\mu}_i^{(\mathrm{VI})}$ and $\boldsymbol{\mu}_i^{(\mathrm{true})}$.

TABLE 1
The Systematic Bias Introduced by Variational Approach for Example 4.2

Dim	Systematic bias by variational approach			
	$\mathrm{dist}(\bar{\boldsymbol{\mu}}_{\mathrm{VI}}, \boldsymbol{\mu}_{\mathrm{true}})$		$\mathrm{dist}(\bar{\lambda}_{\mathrm{VI}}, \lambda_{\mathrm{true}})$	
	$N = 100$	$N = 200$	$N = 100$	$N = 200$
10	0.15	0.08	0.93	0.37
20	0.19	0.10	1.11	0.40
30	0.30	0.14	1.44	0.50
40	0.39	0.19	1.83	0.88
50	0.48	0.27	2.10	1.11

$\boldsymbol{\mu}^{(\mathrm{true})}$ can be obtained by computing the angles between them as:

$$\mathrm{dist}(\bar{\boldsymbol{\mu}}^{(\mathrm{VI})}, \boldsymbol{\mu}^{(\mathrm{true})}) = \sin\theta, \tag{47}$$

where θ is the angle between $\bar{\boldsymbol{\mu}}^{(\mathrm{VI})}$ and $\boldsymbol{\mu}^{(\mathrm{true})}$. Here we have used the angle instead of Euclidean distance, as Euclidean distance is not the natural metrics on the unit sphere (e.g., similar discussions can be found in [36]).

Similarly we can compute the MSE between $\lambda_i^{(\mathrm{VI})}$ and $\lambda^{(\mathrm{true})}$, and $\boldsymbol{\mu}_i^{(\mathrm{VI})}$ and $\boldsymbol{\mu}^{(\mathrm{true})}$ for the ith data set, from which the average MSE is given by:

$$\overline{\mathrm{MSE}}(\lambda_i^{(\mathrm{VI})}, \lambda_i^{(\mathrm{true})}) = \frac{1}{I}\sum_{i=1}^{I}\left(\hat{\lambda}_i^{(\mathrm{VI})} - \lambda^{(\mathrm{true})}\right)^2 \tag{48}$$

$$\overline{\mathrm{MSE}}(\boldsymbol{\mu}_i^{(\mathrm{VI})}, \boldsymbol{\mu}_i^{(\mathrm{true})}) = \frac{1}{I}\sum_{i=1}^{I}\sin^2\theta_i. \tag{49}$$

The result of this experiment is shown in Tables 1 and 2—no experiment with HMC was carried out due to the prohibitive computation time. We observe that the variational approach incurs some bias whose amount becomes smaller with increasing the amount of observed data.

4.2 Log Marginal Likelihood

Here, we experimentally examine the performance of the variational approach in approximating the log marginal likelihood.

Example 4.3. In this synthetic experiment, we generated 100 data points from a complex Watson distribution of the chosen dimensionality $d = \{10, 20, 30, 40, 50\}$. The variational algorithm was terminated after either $1{,}000$ iterations had been reached, or the change in the lower bound on the LML (16) became less than 10^{-6} per datum. To avoid possible local convergence, random parameter initializations were used in an attempt to avoid local maxima—the highest score over random initialisations was taken. For comparison, the Hamiltonian annealed importance sampling (HAIS; [37]) was used to estimate the true LML. HAIS is an extension of the well-known annealed importance sampling (AIS; [38]) for computing normalizing constants, and handling multimodal distributions, using HMC with partial momentum refreshment as the MCMC method. The HAIS was set with 10^5 intermediate distributions and with Hamiltonian dynamics step size of 0.06 and 200 particles. Table 3 illustrates the average LML of data by the variational approach and HAIS. We notice that the LML obtained from the variational posteriors (given by the variational lower bound) assigned similar scores as those based on samples from the true posteriors given by HAIS.

TABLE 2
The Average Mean Square Error Introduced by Variational Approach for Example 4.2

Dim	Average mean square error by variational approach			
	$\overline{\mathrm{MSE}}(\boldsymbol{\mu}_i^{(\mathrm{VI})}, \boldsymbol{\mu}_i^{(\mathrm{true})})$		$\overline{\mathrm{MSE}}(\lambda_i^{(\mathrm{VI})}, \lambda_i^{(\mathrm{true})})$	
	$N = 100$	$N = 200$	$N = 100$	$N = 200$
10	0.022	0.006	0.804	0.140
20	0.035	0.010	1.146	0.163
30	0.088	0.020	1.929	0.255
40	0.149	0.036	3.116	0.792
50	0.226	0.074	4.103	1.2460

TABLE 3
The Average Log Marginal Likelihood Given by Variational Approach and HAIS for Example 4.3

Dim	Average log marginal likelihood (Std err)	
	Variational approach	HAIS
10	−106.63	−104.55 (2.27)
20	−197.64	−194.30 (2.29)
30	−288.62	−284.84 (2.38)
40	−377.35	−372.46 (3.06)
50	−460.40	−453.94 (4.36)

In general, it is difficult to compare computation time of these two algorithms since it would highly depend on choices of stopping criteria in the convergence. Nonetheless, Fig. 4 illustrates the average convergence time over five data sets per dimension. As expected, the variational approach is significantly faster. We also notice that in both cases convergence time per dimension decreases with increasing the dimensionality.

4.3 Model Pruning

The model pruning can be seen as a desirable effect of *Occam's Razor* in Bayesian learning, as the training procedure automatically determines the necessary model complexity and avoids over-fitting [20]. In Bayesian learning of mixture models, this desirable pruning effect is achieved by assigning small values for the parameters $\alpha_{0,k}$ of the prior Dirichlet distribution of component weights $\underline{\tau}$ in (6). With prior values $0 < \alpha_{0,k} \ll 1$, the model favors solutions where

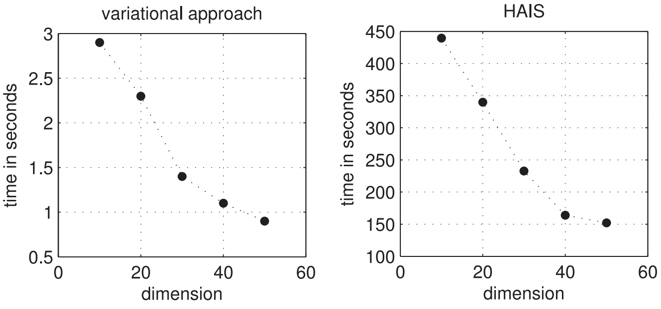


Fig. 4. Average convergence time per dimension for variational approach (left) and HAIS (right) in Example 4.3.

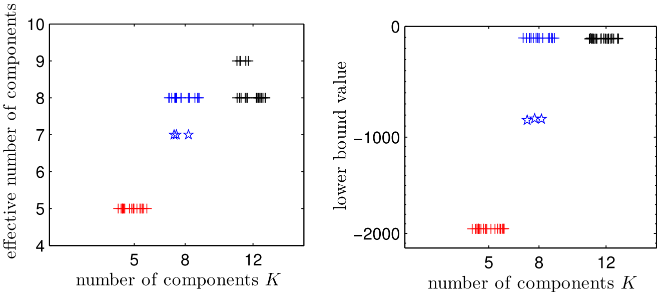


Fig. 5. Model pruning of the mixture components, as evaluated in Example 4.4: (left) the effective number of components remained after training versus the number of components K; (right) the value of the lower bound after convergence versus K. The algorithm is initialized with $K \in \{5, 8, 12\}$ which are indicated respectively with red, blue, and black colors. Cases where results exhibit over-pruning are marked with $/medstar$, whereas cases where there are no indication of over-pruning, the result are marked with +. This figure shows the result of 20 trials where there are random horizontal perturbations for sake of demonstration.

some component weights τ_k approach zero, thus effectively eliminating mixture components that are not really needed to explain the observed data.

In the following, a toy experiment on synthetic data is used to examine the model pruning capability of the method, in the sense of avoiding over-fitting and at the same time avoiding inadequate pruning of the mixture components (over-pruning). The result of this experiment shows that the method performs well in handling model complexity. However, there might be a slight chance of over-pruning, perhaps due to the random initialization of the model parameters.

Example 4.4. For this experiment, a synthetic data set is generated from a known complex Watson mixture model composed of eight components with dimension of $d = 10$ and 8 samples per mixture component on average (we check to validate if there is at least one point generated from each mixture component). The mean direction of each component is a random complex vector on the unit hypersphere. The mixture weights and the concentration parameters are random values sampled from a uniform distribution (the average concentration parameter of all components is about 20). Next, we train different models with the variational method, starting with $K \in \{5, 8, 12\}$ mixture components. Then, for each trained model, we compute the lower bound value after convergence. The experiment is repeated 20 times for the same data set but with different (random) initializations. The result of this experiment is shown in Fig. 5. When we initialize the variational method with too few model components ($K < 8$), the trained model uses all the available model complexity, but the result is still sub-optimal because the model complexity is not sufficient. When we initialize with the exact known number of components ($K = 8$), the training procedure often performs well but sometimes it over-prunes (three times out of 20 trials). This extra pruning is probably due to the random initialization which leads the VI optimization to a sub-optimal local optimum. This indicates that one should perhaps try with several different random initializations, for the same data set, when there is uncertainty whether the assigned model complexity is sufficient. When we initialize with a higher number of components than necessary ($K > 8$), the risk of over-pruning is reduced. For instance, when the initial K was 12, the result sometimes (five times out of 20 times) includes nine components with non-trivial weights, and the training never over-pruned. However, the trained models are all equivalent in relation to the training data, as indicated by the resulting log-likelihood shown in Fig. 5 (right). Thus, we find minor concerns with respect to over-fitting to the data.

5 Application in Source Separation

In this section, we briefly address the problem often referred to as blind source separation. The BSS is the problem of recovering source signals from their mixtures, with only limited knowledge about the source signals and the mixing process. BSS is a central problem in signal processing, appearing in many fields, including speech processing, network tomography and biomedical imaging (e.g., [39]). Here, we only focus on the acoustic applications of BSS, such as solving a cocktail party problem. In many real-world applications, the problem can be seen in under-determined scenarios, that is where there are more signals to estimate (the sources) than signals that are observed (the sensors).

With acoustic applications of BSS, signals are mixed in a convolutive manner with reverberation. Various attempts have been made to solve the convolutive BSS problem. Among them, frequency-domain approaches are of special interest (e.g., [40], [41], [42]), where time domain observation signals are converted into frequency domain time-series signals by a short-time Fourier transform. Many of the existing frequency domain approaches rely on the sparseness of source signals, which means each time-frequency point can be assigned to a single dominant source. A widely used approach to identify the dominant source is to cluster amplitude and phase differences of closely spaced microphone pairs (e.g., [43], [44]), or clustering other time-frequency complex-valued feature vectors extracted from the mixed signals (e.g., [42], [45]).

Recently, Sawada et al. [42] developed a two-stage frequency bin-wise BSS framework which employs a widely used time-frequency masking scheme in the separation stage. In this two-stage approach, the first stage is responsible for frequency bin-wise clustering of time-frequency points—this is the key stage in the separation procedure. In the second stage an additional task is performed to batch bin-wise separated frequency components associated to the same source known as *permutation alignment*. The bin-wise clustering results of the first stage are represented by sets of posterior probabilities which determine the dominant source in each time-frequency point. These posterior probabilities will be again employed for the permutation alignment procedure in the second stage—this is done by exploring similarities among the sequences of posterior probabilities across all frequency bins [46].

The focus of our discussion in this section is on the first stage. We closely follow the frequency bin-wise BSS framework [42]. We show how mixture modeling with complex Watson distributions can fit into this BSS framework, and what potential benefits one can get from formulating the problem in a Bayesian framework with the variational inference.

5.1 System Overview

Let $\{s_i(t)\}_{i=1}^{I}$ denote a set of sources and $\{x_j(t)\}_{j=1}^{J}$ denote a set of observations, where I and J indicate the number of sources and the number of microphones, respectively. Assuming a convolutive mixing model, the observation $x_j(t)$ is given by $x_j(t) = \sum_{i=1}^{I} \sum_l h_{ji}(l) s_i(t-l)$, where $h_{ji}(l)$ represents the impulse response from source i to microphone (sensor) j. The convolutive mixing model $x_j(t)$ is transformed to the time-frequency representation by using the short time Fourier transform, and can be approximated as an instantaneous mixing model at each frequency bin as $x_j(n,f) = \sum_{i=1}^{I} h_{ji}(f) s_i(n,f)$, where $s_i(n,f)$ is the time-frequency representation of $s_i(t)$, f indicates the frequency bin, and n indicates the time frame. In vector notation, it can be written as

$$\mathbf{x}(n,f) = \sum_{i=1}^{I} \mathbf{h}_i(f) s_i(n,f). \tag{50}$$

A set of separated frequency components $\{\widehat{\mathbf{s}}_i(n,f)\}_{i=1}^{I}$, where $\widehat{\mathbf{s}}_i(n,f) = \{\widehat{s}_{i,j}(n,f)\}_{j=1}^{J}$, is obtained by applying a time frequency mask on the observations as

$$\widehat{s}_{i,j}(n,f) = \mathcal{M}_i(n,f) y_j(n,f). \tag{51}$$

Finally time domain sources $\widehat{s}_{i,j}(t)$ are reconstructed by applying an inverse short time Fourier transform to the separated frequency sources $\widehat{s}_{i,j}(n,f)$.

5.2 Clustering Time Frequency Points

Thus the central task is to design proper masks $\mathcal{M}_i(n,f)$. For designing such masks, many approaches rely on the sparseness property of source signals. Under sparsity assumption, (50) can be expressed by

$$\mathbf{x}(n,f) \approx \mathbf{h}_{i^*}(f) s_{i^*}(n,f), \tag{52}$$

where i^* is the index of the dominant source. Hence, we need to identify dominant sources for all time-frequency points. This implies that every observation vector $\mathbf{x}(n,f)$ has to be clustered into I clusters—every cluster corresponds to a certain source. For this purpose, we start with pre-processing of data. We use the standard per-processing procedure as described in [42, Section 3] which includes unit-norm normalization and pre-whitening of data.

5.2.1 Proposed Approach

Let $\underline{\mathbf{x}}_f = \{\mathbf{x}(n,f)\}_{n=1}^{N}$ denote a set of normalized and spatially whitened time-frequency points in the frequency bin f. In this section, we employ the presented approach as a clustering algorithm and model the underlying distribution of $\underline{\mathbf{x}}_f$ by a finite mixture of K complex Watson distributions, as

$$p(\underline{\mathbf{x}}_f \,|\, \underline{\boldsymbol{\mu}}_f, \underline{\lambda}_f, \underline{\tau}_f) = \prod_{n=1}^{N} \sum_{k=1}^{K} \tau_{k,f} \mathcal{W}(\mathbf{x}(n,f) \,|\, \boldsymbol{\mu}_{k,f}, \lambda_{k,f}). \tag{53}$$

The exponential term in (53), $|\boldsymbol{\mu}_{k,f}^{\mathrm{H}} \mathbf{x}(n,f)|^2$, has a particular form which has been suggested, [17], to fit well to the frequency domain BSS task. Firstly, the inner product $\boldsymbol{\mu}_{k,f}^{\mathrm{H}} \mathbf{x}(n,f)$ is analogous to the spatial correlation. The absolute square of this term implies that the response power of the spatially matched beamformer is being used as a distance measure to separate the sources. This concept fits to the beamforming point of view. Secondly, "the morphology of the complex hypersphere reflects spatial aliasing occurring in high frequencies, hence the cyclic nature of the phase differences are implicitly regarded", [17].

It is notable that direction here is not the spatial direction to the sound source. Here, it is the frequency-domain complex gain factors that are modelled by the complex Watson distribution, and these data are "axial" and only indirectly related to the "spatial directions". The fact that the complex Watson distribution is axial does not mean that the "directivity pattern" of the microphone array has equal sensitivity in both axial directions. The complex Watson distribution is invariant to any fixed phase change in all vector elements, but it is still sensitive to the phase differences between vector elements.

We estimate the model parameters using the presented variational approach. Our approach can be considered as an extension to the previous approaches [17], [42], [47], [48], with two main potential advantages which make it possibly preferable in real applications. First, there is no iterative numerical calculation involved in the parameter estimation of the mixture model. In contrast, in [17] the ML estimate to the concentration parameter is obtained by resorting to numerical approaches. However, it should be noted that numerical estimation (e.g., by Newton Raphson methods) of the concentration parameters is nontrivial in high dimensions since it involves ratios of Kummer's hypergeometric function (for more discussion in this regard, readers are referred to [10, Section 3.1]). Second, the presented variational approach can regulate the model complexity by pruning unnecessary mixture components. In [49], this property has been used in order to develop a method for estimating the number of source signals—in other words, relaxing the assumption of prior knowledge on the number of sources in the observations. In contrast, in the EM based approaches, [17], [42], the number of required components in the mixture model is explicitly set to the number of sources, that means a priori knowledge of the number of sources is required.

5.3 Experimental Results and Discussions

We aim to compare the modeling capability of the following two statistical models in modeling frequency observations $\underline{\mathbf{x}}_f = \{\mathbf{x}(n,f)\}_{n=1}^{N}$.

- WMM: mixture of complex Watson distributions.
- GMM: mixture of complex Gaussian distributions projected on the unit hypersphere.

The model parameters in WMM and GMM are estimated by VI and EM, resulting in the following methods:

- VI-WMM—the proposed variational approach;
- EM-WMM—the ML-based approach [10];
- VI-GMM—the variational approach [48];
- EM-GMM—the ML-based approach [42].

To have a fair comparison, we assume the number of sources (mixture components) is given as prior knowledge in all cases, in this experiment.

Evaluations are performed on SiSEC08 [50] database (Development Data). We consider two-channel live

TABLE 4
Comparison of the Modeling Capability of $\mathrm{Model\ A}$ and $\mathrm{Model\ B}$ in Modeling the Time Frequency Observations for a Given Mixture in Various Conditions ($N = 240$)

	Average log marginal likelihood $\mathfrak{L}$ (Std err)			
mix	VI-WMM	EM-WMM	VI-GMMM	EM-GMM
A	−440.21 (3.13)	−**431.21** (2.21)	−496.30 (2.80)	−512.72 (4.23)
B	−511.14 (3.39)	−**501.33** (2.60)	−571.93 (3.03)	−585.95 (5.92)
C	−**589.01** (3.78)	−600.11 (4.89)	−693.39 (3.25)	−708.14 (6.13)
D	−**702.82** (4.41)	−712.82 (6.10)	−845.76 (3.97)	−853.52 (7.45)

The table illustrates the average log marginal likelihood and standard error across five trials.

recordings of 5 seconds female speech signals sampled at 16 kHz for various conditions. For convenience, let triple $(\mathrm{I}, \mathrm{RT}, \mathrm{L})$ denote the number of sources, reverberation time, and microphone spacing, respectively. We consider the following scenarios:

- mix. A: the data has $(\mathrm{I}, \mathrm{RT}, \mathrm{L}){=}(3, 0.13\ \mathrm{s}, 1\ \mathrm{m})$;
- mix. B: the data has $(\mathrm{I}, \mathrm{RT}, \mathrm{L}){=}(3, 0.25\ \mathrm{s}, 1\ \mathrm{m})$;
- mix. C: the data has $(\mathrm{I}, \mathrm{RT}, \mathrm{L}){=}(3, 0.13\ \mathrm{s}, 5\ \mathrm{cm})$;
- mix. D: the data has $(\mathrm{I}, \mathrm{RT}, \mathrm{L}){=}(3, 0.25\ \mathrm{s}, 5\ \mathrm{cm})$.

The model comparison should be performed based on the true LML. Since the true LML is not available, we consider the estimated LML. For EM-based approaches (i.e., EM-WMM and EM-GMM), the LML is the expectation of the data likelihood under a Dirac function about the ML estimates.[7] For VI-based approaches (i.e., VI-WMM and VI-GMM), the approximate LML is given by the lower bound which is the expected likelihood under the posterior distribution over parameters.[8]

Let $\mathfrak{L}_f$ denote the estimated LML for the frequency observation $\underline{\mathbf{x}}_f$ given the model in frequency bin f. Then, we define the average LML of the whole data (i.e., including all frequency bins) as $\mathfrak{L} = \frac{1}{F}\sum_{f=1}^{F} \mathfrak{L}_f$, where F is the number of frequency bins. Table 4 illustrates results of the experiment over five trials. We notice that modeling with Watson distributions (WMM) assigned higher likelihood compared to the popular modeling with bounded Gaussian distributions (GMM), that is both VI-WMM and EM-WMM scored higher. The result of the experiment suggests that WMM might be a better candidate than GMM for modeling time-frequency feature vectors extracted from the mixed signals.

Now the question is how much gain one can get from the further modeling capability offered by modeling with Watson distributions. To answer this question, we adopt the framework by Sawada et al. [42], and look at the corresponding separation results in terms of the output signal-to-distortion ratio (SDR) [51]. Table 5 shows average results of five trials for all sources—we recall that the number of components (sources) is assumed to be known for all algorithms for a fair comparison. Surprisingly, we notice that *the WMM has only slightly better performance compared to the GMM* despite considerably better modeling capability, as measured by LML. One explanation is that the final separation performance also depends on the permutation alignment procedure—a procedure for solving the well-known *permutation ambiguity* problem by reordering the indices of separated clusters in each frequency bin so that the same index corresponds to the same source over all frequencies [42]. To validate this, in the following, we design a toy experiment in which we assume to have an "ideal permutation".[9]

TABLE 5
Separation Results in Terms of the Average Output Signal-to-Distortion Ratio of All Sources

	Average signal-to-distortion ratio in dB			
mix	VI-WMM	EM-WMM	VI-GMM	EM-GMM
A	7.21 (0.69)	**7.33** (0.59)	6.85 (0.69)	7.12 (0.82)
B	6.77 (0.49)	**6.85** (0.41)	6.28 (0.85)	6.68 (0.61)
C	**6.24** (0.47)	6.12 (0.53)	5.74 (0.66)	5.92 (0.53)
D	**5.53** (0.43)	5.33 (0.63)	5.08 (0.45)	5.27 (0.40)

Example 5.1. We generate a two-channel synthetic recording of 6 seconds of three speech signals sampled at 8 kHz for a given reverberation time varying from 100 to 500 ms—the room impulse responses for the experiments are obtained by using Lehmann and Johansson's implementation of the image-source method [52]. We only consider VI-WMM and EM-GMM as representatives of WMM and GMM. Fig. 6 illustrates the results. First we observe that, as expected, the ideal permutation resulted in higher SDR values for both algorithms—more noticeable for high reverberation times. However, more importantly, we notice that the "ideal permutation" for EM-GMM only resulted in small improvement in separation performance, while for VI-WMM, the improvement is more distinguishable.

To summarize, our experiments show that the current frequency domain BSS approaches may potentially benefit from mixture modeling with Watson distributions (WMM). However, the benefit can be realized only with reliable permutation alignment strategies. We emphasize the necessity of further effort toward exploring new permutation alignment strategies or developing permutation-free BSS approaches.

As discussed earlier, owing to the desirable property of model pruning, the variational method can regulate the model complexity and avoid overfitting. In [49], we have described a method for estimating the number of sources using this desirable property.

6 Application in Gene Expression Data

An important step in the analysis of gene expression data is the clustering of genes into groups exhibiting similar expression values over a range of independent experiments. Given statistically sufficient independent experiments, genes grouped in the same cluster share the same functional

[7] This is an upper bound only if ML estimates are the global maximum of the likelihood. This may not happen due to local maxima in the optimization process due to the EM-type optimization being employed.

[8] This can be proved by a convexity bound to be strictly upper bound on the true marginal likelihood.

[9] Under the frequency bin-wise BSS framework, there is always permutation ambiguity. What we consider here as the "ideal permutation" was obtained through exhaustive search considering all the possible pair-wise frequencies with no time constraint in the permutation alignment algorithm by Sawada et al., [42]. However, it should be noted that this still cannot be regarded as a truly permutation free method.

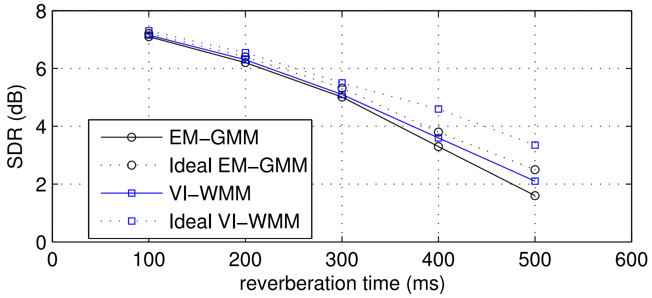


Fig. 6. Average SDR for the given reverberation time. The methods with the ideal permutation alignment procedure are shown by Ideal EM-GMM and Ideal VI-VMM.

behaviour (for more discussions in this regard, readers are referred to [53]). Generally speaking, clustering methods differ from each other based on choice of distortion measure or in other word the measure of similarity.

Clustering using Pearson correlation (as the measure of similarity) is fairly common for the analysis of gene expression data (e.g., [4], [54], [55]), motivated by the hypothesis that for such data relative spatial orientations of the data sample vectors are more important than the magnitudes of the data sample vectors. It has been also noted that the clustering of data using Pearson correlation is essentially a clustering problem for directional data [4]. For instance, Banerjee et al. [4] and Taghia et al. [56] employed mixture modeling with the von Mises-Fisher distributions for this application domain. Their result show significant improvement in clustering performance in comparison to the conventional approaches, e.g., frequency sensitive spherical K-means [55] and the spherical K-means [54].

All these algorithms tend to put those genes in one cluster that show strong positive correlation between their expression vectors. However, as discussed by Shatkay et al., [12], "Genes that are functionally related may demonstrate strong anti-correlation in their expression levels, and thus be clustered into separate groups, blurring the (functional) relationship between them". Based on these observations, Dhillon et al., [11], proposed a clustering algorithm which puts strongly correlated and anti-correlated genes into the same "diametric" cluster. The clustering method is known as diametrical clustering which has some resemblance to the K-means clustering method.

The observation by Shatkay et al., [12] might actually suggest that gene expression data be treated as axial data rather than directional data. Sra and Karp [10] treated gene expression data as axial data. They developed a clustering method based on mixture modeling using Watson distributions. They further showed that the diametrical clustering can be regarded as a special case of mixture-modelling with Watson distributions such that the additional modeling power can lead to a better clustering.

6.1 Experimental Set-Up

In the following, we regard gene-expression data once as directional data and cluster them using mixture of vMF distributions [56]. Next, we regard gene-expression data as axial data and cluster them using mixture of Watson distributions.

TABLE 6
Clustering Accuracy on Gene-Expression Data Set Over 10 Trials

Data	Average homogeneity and separation			
	VI-WMM		VI-vMFMM	
	$\mathrm{H}_{\mathrm{ave}}$	$\mathrm{S}_{\mathrm{ave}}$	$\mathrm{H}_{\mathrm{ave}}$	$\mathrm{S}_{\mathrm{ave}}$
A	+0.56 (0.04)	−**0.71** (0.03)	+0.54 (0.07)	−0.13 (0.02)
B	+0.38 (0.06)	−**0.29** (0.04)	+0.40 (0.09)	−0.03 (0.01)
C	+**0.46** (0.05)	−**0.44** (0.03)	+0.43 (0.04)	−0.05 (0.02)
D	+0.56 (0.06)	−**0.65** (0.01)	+**0.61** (0.13)	−0.06 (0.03)

For $\mathrm{H}_{\mathrm{ave}}$ higher values and for $\mathrm{S}_{\mathrm{ave}}$ lower values are preferable. Noticeable differences (i.e., > 0.02) are shown in bold. The table illustrates the average and standard error across 10 trials.

For this experiment, we consider four gene microarray data sets[10]:

- Data A: Diauxic Shift [57], with matrix size of 826×7;
- Data B: Yeast Cell Cycle [58], with matrix size of $1{,}000 \times 23$ (these 1,000 genes were randomly selected from the original 4,382 genes);[11]
- Data C: Human Fibroblasts [59], with matrix size of 517×18;
- Data D: Sporulation in Yeast [60], with matrix size of $1{,}000 \times 7$ (these 1,000 genes were randomly selected from the original 6,118 genes).

Data sets are normalized to have unit norm. The unit norm normalization does not affect the concept of Pearson correlation [4].

6.2 Experimental Results

In this application domain we do not have the ground-truth clusterings, which can further motivate using the presented approach. We compare clustering results of mixture modeling with Watson distributions with those of with vMF distributions [56] in the Bayesian framework. Since the true cluster labels are not available, the clustering performance is often evaluated by computing certain internal figures of merit. Measuring the homogeneity and the separation of the clusters are two commonly used figures of merit (e.g., [4], [61]). The homogeneity determines how similar the individual members of each cluster are to their cluster representative, hence, it provides a measure of the intra-cluster similarity (higher values are favorable). The separation determines how disjoint the clusters are from each other, hence, it provides a measure of inter-cluster similarity (lower values are favorable). Let $\underline{\mathbf{X}} = \{\mathbf{x}_n\}_{n=1}^{N}$ denote a set of observed data and $\{\Omega_j\}_{j=1}^{J}$ denote disjoint clusters where Ω_j includes $\mathbf{x}_n$ which belong to the jth cluster. As defined in [10], the average homogeneity $\mathrm{H}_{\mathrm{ave}}$ and separation $\mathrm{S}_{\mathrm{ave}}$ measures can be evaluated as

$$\mathrm{H}_{\mathrm{ave}} = \frac{1}{N} \sum_{j=1}^{J} \sum_{\mathbf{x} \in \Omega_j} \left(\mathbf{x}^{\mathrm{T}} \boldsymbol{\mu}_j\right)^2, \tag{54}$$

$$\mathrm{S}_{\mathrm{ave}} = \frac{1}{\sum_{i \neq j} |\Omega_i||\Omega_j|} \sum_{i \neq j} |\Omega_i||\Omega_j| \min\left(\boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{\mu}_j, -\boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{\mu}_j\right), \tag{55}$$

[10] Download: http://transcriptome.ens.fr/gepas/data/index.html
[11] www.exploredata.net/Downloads/Gene-Expression-Data-Set

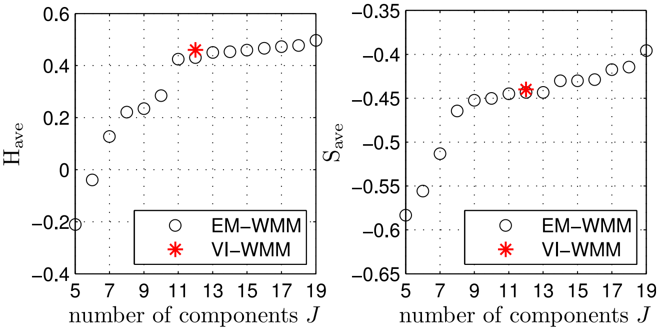


Fig. 7. Measures of cluster quality on Human Fibroblasts data (Data C): (left) the average homogeneity H_{ave}; (right) the average separation S_{ave}, across 10 trials. For H_{ave} higher values and for S_{ave} lower values are preferable.

where $|\Omega_i|$ determines the number of members of the ith cluster. Note that $J \in \{1, \ldots, K\}$ is the number of clusters which is determined automatically during modeling, and K is the initial guess of the number of clusters which is set to a large value, $K > J$. We note the slight departure from the standard definitions in (54) and (55).[12] Table 6 illustrates the results. We notice that interestingly mixture modeling with Watson distributions (shown by VI-WMM in the table) compared to mixture modeling with vMF distributions (shown by VI-vMFMM in the table) yields clusters having similar intra-cluster cohesiveness but significantly better inter-cluster separation. One possible reason is that VI-WMM resulted in fewer clusters.

Our results suggest that the gene-expression data may be treated as axial data rather than directional data, and may statistically be modeled using axial distributions (e.g., the Watson distribution) rather than directional distributions (e.g., the vMF distribution).

We recall that the variational approach (VI-WMM) regulates its own complexity by determining necessary components. This is an interesting property that makes our approach more desirable compared to its EM formulation (EM-WMM) [10]. Fig. 7 compares the clustering results of VI-WMM and EM-WMM on Human Fibroblasts data set (Data C). For the EM-WMM, the number of clusters J varies from 5 to 19, while for the VI-WMM, it is determined automatically during optimization. This figure shows that for the EM-WMM, by increasing cluster numbers, the intra-cluster similarity improves but the inter-cluster similarity degrades. Hence, there is a compromise involved so that one may need several runs of the EM-WMM in order to determine the right number of clusters satisfying this compromise. However, the VI-WMM can find nearly the right number of components over a single run, and hence preferable in practice.

7 Conclusion

We developed a variational inference algorithm for Watson mixture model and demonstrated its applicability as an alternative to computationally costly MCMC sampling methods. While MCMC sampling can provide theoretical guarantees of accuracy, the variational method provides a fast, deterministic approximation to the posterior distribution. We briefly addressed the potential application of mixture modeling with complex Watson distributions in the problem of blind source separation and with real Watson distributions in clustering gene microarray data.

Acknowledgments

This work was funded by the European Commission within the Marie Curie ITN AUDIS, grant PITNGA-2008-214699.

References

[1] K. V. Mardia and P. E. Jupp, *Directional Statistics*. Hoboken, NJ, USA: Wiley, 2000.

[2] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 25, pp. 14 863–14 868, 1998.

[3] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. Int. Conf. World Wide Web*, 2001, pp. 285–295.

[4] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *J. Mach. Learning Res.*, vol. 6, pp. 1345–1382, 2005.

[5] A. Strehl, E. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," in *Proc. Workshop Artif. Intell. Web Search.*, 2000, pp. 58–64.

[6] G. S. Watson, "Equatorial distributions on a sphere," *Biometrika*, vol. 52, pp. 193–201, 1965.

[7] K. V. Mardia and I. L. Dryden, "The complex Watson distribution and shape analysis," *J. Roy. Statistical Soc. Series B (Statistical Methodol.)*, vol. 61, no. 4, pp. 931–926, 1999.

[8] I. L. Dryden, "Statistical analysis on high-dimensional spheres and shape spaces," *Anna. Statist.*, vol. 33, no. 4, pp. 1643–1665, 2005.

[9] A. S. Bijral, M. Breitenbach, and G. Grudic, "Mixture of Watson distributions: A generative model for hyperspherical embeddings," in *Proc. 11th Int. Conf. Artif. Intell. Statist.*, 2007, pp. 35–42.

[10] S. Sra and D. Karp, "The multivariate Watson distribution: Maximum-likelihood estimation and other aspects," *J. Multivariate Anal.*, vol. 114, pp. 256–269, 2011.

[11] I. S. Dhillona, E. M. Marcotte, and U. Roshan, "Diametrical clustering for identifying anti-correlated gene clusters," *Bioinformatics*, vol. 19, pp. 1612–1619, 2003.

[12] H. Shatkay, S. Edwards, W. J. Wilbur, M. Boguski, W. John, and W. M. Boguski, "Genes, themes and microarrays: Using information retrieval for large-scale gene analysis," in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 2000, pp. 317–328.

[13] K. P. Lennox, D. B. Dahl, and M. A. Vannucci, "Density estimation for protein conformation angles using a bivariate von Mises distribution and Bayesian nonparametrics," *J. Am. Statist. Assoc.*, vol. 104, pp. 586–596, 2009.

[14] K. V. Mardia, C. C. Taylor, and G. K. Subramaniam, "Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data," *Bioinformatics*, vol. 63, pp. 505–512, 2007.

[15] A. Bhattacharya and D. Dunson, "Nonparametric Bayesian density estimation on manifolds with applications to planar shapes," *Biometrika*, vol. 97, no. 4, pp. 851–865, 2010.

[16] P. Common and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, 1st ed. Norwell, MA, USA: Academic, 2010.

[17] D. H. Tran-Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2010, pp. 241–244.

[18] O. Besson and S. Bidon, "Robust adaptive beamforming using a Bayesian steering vector error model," *Signal Process.*, vol. 93, pp. 3290–3299, 2013.

[19] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York, NY, USA: Springer-Verlag, 2004.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[21] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.

[12] In standard definition, explicitly in the case of the vMF distribution, in (54) instead of summing over $\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\mu}_j\right)^2$, we take the sum over $\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\mu}_j\right)$; and in (55) instead of $\min\left(\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_j, -\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_j\right)$, we only have $\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_j$.

[22] G. Andrews, R. Askey, and R. Roy, *Special Functions*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[23] A. B. O. Daalhuis, *The NIST Handbook of Mathematical Functions*. F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[24] M. J. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.

[25] R. E. Turner and M. Sahani, "Two problems with variational expectation maximisation for time-series models," in *Bayesian Time Series Models*, D. Barber, T. Cemgil, and S. Chiappa, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2011, ch. 5, pp. 109–130.

[26] Z. Ghahramani and M. J. Beal, "Propagation algorithms for variational Bayesian learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 507–513.

[27] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *Anna. Appl. Statist.*, vol. 1, pp. 17–35, 2007.

[28] M. Braun and J. McAuliffe, "Variational inference for large-scale models of discrete choice," *J. Am. Statistical Assoc.*, vol. 105, pp. 324–335, 2007.

[29] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2160–2173, Nov. 2011.

[30] M. J. Prentice, "A distribution-free method of interval estimation for unsigned directional data," *Biometrika*, vol. 71, no. 1, pp. 147–154, 1984.

[31] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[32] R. Salakhutdinov and S. Roweis, "Adaptive overrelaxed bound optimization methods," in *Proc. Int. Conf. Mach. Learn.*, 2003, vol. 20, pp. 664–671.

[33] R. Turner, P. Berkes, M. Sahani, and D. Mackay. (2008). Counterexamples to variational free energy compactness folk theorems [Online]. Available: http://www.gatsby.ucl.ac.uk/ turner/Notes/ Compactness/ CompactnessFolkTheorem.pdf, Tech. Rep.

[34] R. Neal, "MCMC using Hamiltonian dynamics," in S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds. *Handbook of Markov Chain Monte Carlo.* Boca Raton, FL, USA: CRC Press, 2011.

[35] P. D. Hoff, "Simulation of the matrix Bingham von Mises-Fisher distribution, with applications to multivariate and relational data," *J. Comput. Graph. Statist.*, vol. 18, no. 2, pp. 438–456, 2009.

[36] O. Besson, N. Dobigeon, and J.-Y. Tourneret, "CS decomposition based Bayesian subspace estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4210–4218, Aug. 2012.

[37] J. Sohl-Dickstein and B. Culpepper, "Hamiltonian annealed importance sampling for partition function estimation," *arXiv preprint arXiv:1205.1925*, p. 10, 2012. [Online]. Available: http://arxiv.org/abs/1205.1925

[38] R. Neal, "Annealed importance sampling," *Statist. Comput.*, vol. 11, no. 2, pp. 125–139, 2001.

[39] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Hoboken, NJ, USA: Wiley, 2001.

[40] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, nos. 1–3, pp. 21–34, 1998.

[41] S. Makino, T. Lee, and H. Sawada, *Blind Speech Separation*. New York, NY, USA: Springer, 2007.

[42] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech Language Process.*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[43] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[44] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.

[45] P. D. O'Grady and B. A. Pearlmutter, "The LOST algorithm: Finding lines and separating speech mixtures," *EURASIP J. A. Signal Process.*, vol. 2008, p. 784296, 2008.

[46] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain bss," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2007, pp. 3247–3250.

[47] D. H. Tran-Vu and R. Haeb-Umbach, "An EM approach to integrated multichannel speech separation and noise suppression," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2010.

[48] J. Taghia, N. Mohammadiha, and A. Leijon, "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 253–256.

[49] J. Taghia and A. Leijon, "Separation of unknown number of sources," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 625–629, May 2014.

[50] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. 8th Int. Conf. Independent Compon. Anal.*, 2009, pp. 253–256.

[51] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. 7th Int. Conf. Independent Compon. Anal. Signal Separation*, 2007, pp. 552–559.

[52] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 269–277, 2008.

[53] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 25, pp. 14 863–14 868, Dec. 1998.

[54] I. S. Dhillon, J. Fan, and Y. Guan, *Efficient Clustering of Very Large Document Collections*. Norwell, MA, USA: Kluwer, 2001, pp. 357–381.

[55] A. Banerjee and J. Ghosh, "Frequency sensitive competitive learning for clustering on high-dimensional hyperspheres," in *Proc. Int. Joint Conf. Neural Netw.*, May 2002, pp. 1590–1595.

[56] J. Taghia, Z. Ma, and A. Leijon, "Bayesian estimation of the Von-Mises Fisher mixture model with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 9, pp. 1701–1715, Sep. 2014.

[57] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997.

[58] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R, K. Anders, M. B. Eisen, P. O. Brown, B. Futcher, and G. R. Fink, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Molecular Biol. Cell*, vol. 9, pp. 3273–3297, 1998.

[59] V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. Hudson, M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P. O. Brown, "The transcriptional program in the response of human fibroblasts to serum," *Science*, vol. 283, no. 5398, pp. 83–87, 1999.

[60] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz, "The transcriptional program of sporulation in budding yeast," *Science*, vol. 282, no. 5389, pp. 699–705, 1998.

[61] R. Sharan and R. Shamir, "CLICK: A clustering algorithm with applications to gene expression analysis," in *Proc. Int. Conf. Intell. Syst. Molecular Biol.*, 2000, pp. 307–316.

Jalil Taghia received the PhD degree in electrical engineering from KTH Royal Institute of Technology in 2014. He is a postdoctoral researcher at the Cognitive and Systems Neuroscience Laboratory at Stanford University. Since June 2014 till August 2015, he was a postdoctoral researcher at Neural Information Processing Group, Technical University of Berlin.

Arne Leijon received the MSc degree (Civilingenjör) in engineering physics in 1971, and the PhD degree in information theory in 1989, both from Chalmers University of Technology in Gothenburg, Sweden. He is a professor em. in hearing technology at KTH Royal Institute of Technology, Stockholm, Sweden. His main research interest concerns applied signal processing in aids for people with hearing impairment, and methods for individual fitting of these devices.