Contents

1

A simple and general exponential family framework for partial membership and factor analysis

Shakir Mohamed

Department of Computer Science, University of British Columbia

Katherine A. Heller

Department of Statistical Science, Duke University

Zoubin Ghahramani

Department of Engineering, University of Cambridge

CONTENTS

We show how mixture models, partial membership models, factor analysis, and their extensions to more general mixed-membership models, can be unified under a simple framework using the exponential family of distributions and variations in the prior assumptions on the latent variables that are used. We describe two models within this common latent variable framework: a Bayesian partial membership model and a Bayesian exponential family factor analysis model. Accurate inferences can be achieved within this framework that allow for prediction, missing value imputation, and data visualisation, and importantly, allow us to make a broad range of insightful probabilistic queries of our data. We emphasise the adaptability and flexibility of these models for a wide range of tasks, characteristics that will continue to see such models used at the core of modern data analysis paradigms.

1.1 Introduction

Latent variable models are ubiquitous in machine learning and statistics and are core components of many of the most widely used probabilistic models, including mixture models [39, 9], factor analysis [3], probabilistic principal components analysis [47, 9], mixed-membership models [15], and matrix factorisation [29, 43], amongst others. The use and success of latent variables lies in that they provide us a mechanism with which to achieve many of the desiderata of modern data modelling: robustness to noise, allowing for accurate predictions of future events, the ability to handle and impute missing data, and providing insights into the phenomena underlying our data. For example, in mixture models the latent variables represent the membership of data points to one of a set of underlying classes, or in topic models, the latent variables allow us to represent the distribution of topics captured within a set of documents.

The broad applicability of mixed-membership models is expanded upon throughout this volume, and here we shall focus on simpler instances of the general mixed-membership modelling framework to emphasis this wide-applicability. In this chapter, we show how mixture models, factor analysis and partial membership models, and their generalisation to mixed-membership models can be unified under a common modelling framework. Moreover, we show how exponential family likelihoods can be used to provide a very general tool for modelling diverse data types, such as binary, count or non-negative data, etc. Specifically, we will develop two models: A Bayesian partial membership model (BPM) [23] and a Bayesian exponential family factor analysis (EXFA) [35], and demonstrate the power of these models for accurate prediction and interpretation of data.

We will use as a case study, the analysis of recorded votes: data that lists the names of those voting for or against a motion. In particular, we will focus on the roll call of the US senate, and demonstrate the different perspectives of our data that can be obtained, and the types of probabilistic queries that can be made with an accurate model of the data. Recorded votes are stored as a binary matrix and we describe a general approach for handling this type of data, and generally, any data that can be described by members of the exponential family of distributions. We develop two probabilistic models: the first is a model for *partial memberships* that allows us to describe senators on a scale of fully-allegiant Democrats to fully-allegiant Republicans. This is a natural way of thinking of such data, since senators are often grouped into blocs depending on their degree of membership to these two groups, such as moderate Democrats, Republican majority, etc. Secondly, we develop *factor models* that provide a means of representing the underlying factors or traits that senators use in their decision making. These two models will be shown to arise naturally from the same probabilistic framework, allowing us to explore different assumptions on the underlying structure of the data.

We begin our exposition by providing the required background on conjugate-exponential family models (section 1.2.1). We then show that by considering a relaxation of standard mixture models we arrive naturally at two useful model classes: latent Dirichlet models and latent Gaussian models (which we expand upon in section 1.4). In section 1.2.3, we show that the assumption of Dirichlet distributed latent variables allows us to develop a model that quantifies the partial membership of objects to clusters, and that the assumption of continuous, unconstrained latent variables in section 1.2.4 leads to an exponential family factor analysis. We focus on Markov chain Monte Carlo methods for learning in both models in section 1.3. Whereas many types of mixed-membership models focus on representing the data at two levels (e.g., a subject and a population level), here we operate at one level (subject level) only, and we describe the relationship between our approach and other mixed-membership models such as Latent Dirichlet Allocation and mixed-membership matrix factorisation [10, 15, 32] in section 1.4. We provide some experimental results and explore the roll call data in section 1.5.

Notation: Throughout this chapter we represent observed data as an $N \times D$ matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$, with an individual data point $\mathbf{x}_n = [x_{n1}, \ldots, x_{nD}]$. N is the number of data points and D is the number of input features. $\boldsymbol{\Theta}$ is a $K \times D$ matrix of model parameters with rows $\boldsymbol{\theta}_k$. $\mathbf{V}$ is a $N \times K$ matrix $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]^\top$ of latent variables with rows $\mathbf{v}_n = [v_{n1}, \ldots, v_{nK}]$, which are K-dimensional vectors of continuous values in $\mathbb{R}$. K is the number of latent factors representing the dimensionality of the latent variable.

1.2 Membership Models for the Exponential Family

1.2.1 The Exponential Family of Distributions

The choice of likelihood function $p(\mathbf{x}|\boldsymbol{\eta})$, for parameters $\boldsymbol{\eta}$ and observed data $\mathbf{x}$, is central to the models we describe here. In particular, we would like to model data of different types, i.e. data that may be binary, categorical, real-valued, etc. To achieve this objective, we make use of the exponential family of distributions, which is an important family of distributions that emphasises the shared properties of many standard distributions, including the Binomial, Poisson, Gamma, Beta, Multinomial, and Gaussian distributions [7]. The exponential family of distributions allows us to provide a singular discussion of the inferential properties associated with members of the family and thus, to develop a modelling framework generalised to all members of the family.

In the exponential family of distributions, the conditional probability of $\mathbf{x}$ given parameter value $\boldsymbol{\eta}$ takes the following form:

$$p(\mathbf{x}|\boldsymbol{\eta}) = \exp\{s(\mathbf{x}_n)^\top \boldsymbol{\eta} + h(\mathbf{x}_n) - g(\boldsymbol{\eta})\}, \tag{1.1}$$

where $s(\mathbf{x}_n)$ are the sufficient statistics, $\boldsymbol{\eta}$ is a vector of natural parameters, $h(\mathbf{x}_n)$ is a function of the data and $g(\boldsymbol{\eta})$ is the cumulant or log-partition function. For this chapter, the natural representation of the exponential family likelihood is used, such that $s(\mathbf{x}) = \mathbf{x}$. For convenience, we shall represent a variable $\mathbf{x}$ that is drawn from an exponential family distribution using the notation: $\mathbf{x} \sim \text{Expon}(\boldsymbol{\eta})$, with natural parameters $\boldsymbol{\eta}$.

Probability distributions that belong to the exponential family also have corresponding conjugate prior distributions $p(\boldsymbol{\eta})$, for which both $p(\boldsymbol{\eta})$ and $p(\mathbf{x}|\boldsymbol{\eta})$ have the same functional form. The conjugate prior distribution for the exponential family distribution of equation (1.1) is:

$$p(\boldsymbol{\eta}) \propto \exp\{\boldsymbol{\lambda}^\top \boldsymbol{\eta} - \nu g(\boldsymbol{\eta}) + f(\boldsymbol{\lambda})\}, \tag{1.2}$$

where $\boldsymbol{\lambda}$ and ν are hyperparameters of the prior distribution. We use the shorthand: $\boldsymbol{\eta} \sim \text{Conj}(\boldsymbol{\lambda}, \nu)$ to denote draws from a conjugate distribution.

As an example, consider binary data, for which an appropriate data distribution is the Bernoulli distribution and the corresponding conjugate prior is the Beta distribution. The Bernoulli distribution has the form: $p(x|\mu) = \mu^x(1-\mu)^{1-x}$, with μ in [0,1]. The exponential family form, using the terms in equation (1.1), is described using: $h(x) = 0$, $\eta = \ln(\frac{\mu}{1-\mu})$ and $g(\eta) = \ln(1+e^\eta)$. The natural parameters can be mapped to the parameter values of the distribution using the link function, which is the logistic sigmoid in the case of the Bernoulli distribution. The terms of the conjugate distribution can also be derived easily.

1.2.2 Beyond Mixture Models

Mixture models are a common approach for assigning membership of observations to a set of distinct clusters. For a finite mixture model with K mixture components, the probability of a data observation $\mathbf{x}_n$ given parameters $\boldsymbol{\Theta}$ is:

$$p(\mathbf{x}_n|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \rho_k p_k(\mathbf{x}_n|\boldsymbol{\theta}_k), \tag{1.3}$$

where $p_k(\cdot)$ is the probability distribution of mixture component k, and ρ_k is the mixing proportion. We can express this using indicator variables $\mathbf{v}_n = [v_{n1}, v_{n2}, \ldots, v_{nK}]$ as:

$$p(\mathbf{x}_n|\boldsymbol{\Theta}) = \sum_{\mathbf{v}_n} p(\mathbf{v}_n) \prod_{k=1}^{K} \left(p_k(\mathbf{x}_n|\boldsymbol{\Theta}_k)\right)^{v_{nk}}, \tag{1.4}$$

where $v_{nk} \in \{0, 1\}$, $\sum_k v_{nk} = 1$, and $p(v_{nk} = 1) = \rho_k$. If $v_{nk} = 1$ then observation n belongs to cluster k, and therefore v_{nk} indicates the membership of observations to clusters.

We now consider a relaxation of this model: relaxing the constraint that $v_{nk} \in \{0, 1\}$ to instead be continuous-valued and removing the sum-to-one constraint. The probability in equation 1.4 must now be modified, and becomes:

$$p(\mathbf{x}_n|\boldsymbol{\Theta}) = \int_{\mathbf{v}_n} p(\mathbf{v}_n) \frac{1}{Z(v_n, \boldsymbol{\Theta})} \prod_{k=1}^{K} \left(p_k(\mathbf{x}_n|\boldsymbol{\Theta}_k)\right)^{v_{nk}} d\mathbf{v}_n, \tag{1.5}$$

where we have integrated over the continuous latent variables rather than summing, and have introduced the normalising constant Z, which is a function of $\mathbf{v}_n$ and $\boldsymbol{\Theta}$ to ensure normalisation.

By substituting the exponential family distribution 1.1 into equation 1.5, the likelihood can be expressed as:

$$\mathbf{x}_n|\mathbf{v}_n, \boldsymbol{\Theta} \sim \text{Expon}\left(\sum_k v_{nk}\boldsymbol{\theta}_k\right), \tag{1.6}$$

which is obtained by combining terms in log-space and requiring the resulting distribution to be normalised. The computation of the normalising constant Z in equation 1.5 is thus always tractable. Thus, we see that the observed data can be described by an exponential family distribution with natural parameters that are given by the linear combination of the coefficients $\boldsymbol{\theta}_k$ weighted by the latent variables v_{nk}.

We consider two types of constraints on the latent variables, which give rise to two important model classes. These are:

TABLE 1.1
Models which can be derived from the unifying framework for latent variable models.

Model	**Domain**
Mixture Models	$v_{nk} \in \{0, 1\}$
Partial Membership Models [23]	$v_{nk} \in [0, 1]$
Exponential Family PCA [13, 35]	$v_{nk} \in \mathbb{R}$
Non-negative Matrix Factorisation [29]	$v_{nk} \in \mathbb{R}^+$

Partial-membership models. The latent variables can take any value in the range $v_{nk} \in [0, 1]$. It is with this relaxation that we are able to represent data points that can belong partially to a cluster. Such ideas are found in fuzzy set theory, and mixed-membership and topic modelling.

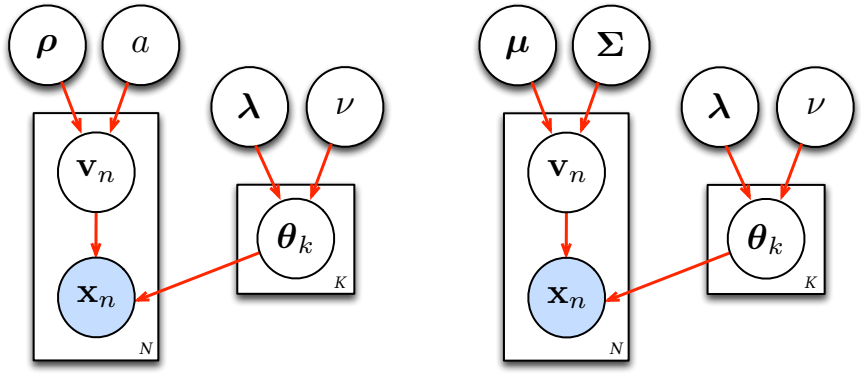
Factor models. The latent variables are allowed to take any continuous value $v_{nk} \in \mathbb{R}$. Popular models that stem from this assumption include factor analysis (FA) [3], probabilistic principal components analysis (PCA) [47], and probabilistic matrix factorisation (PMF) [43], amongst others. The latent variables form a continuous low-dimensional representation of the input data. For easier interpretation, one can restrict the latent variables to be non-negative, allowing for a parts based explanation of the data [29].

Thus, we obtain a unifying framework for many popular latent variable models, whose key difference lies in the nature of the latent variables used. Table 1.1 summarises this insight and lists some of the models that can arise from this framework.

1.2.3 Bayesian Partial Membership Models

We consider a model for partial membership that we refer to as the Bayesian Partial Membership model (BPM) [23]. The BPM is a model in which we consider observations that exist in a space of K classes, but in which an observation has partial membership all the available classes. Consider political affiliations as an example, an individual's political leaning is not wholly socialist or wholly conservative, but may have partial membership in both these political schools.

At the outset it is important to note the distinction between partial membership and uncertain membership. Responsibilities in mixture models are representations of the uncertainty in assigning full membership to a cluster, and this uncertainty can often be reduced with more data. Partial membership represents a fractional membership in multiple clusters, such as a senator with moderate views inbetween that of being fully Republican or fully Democrat.



(a) Bayesian Partial Membership

(b) Exponential Family Factor Analysis

FIGURE 1.1
Graphical models representing the relationship between latent variables, parameters and observed data for exponential family latent variable models.

Figure 1.1(a) is a graphical representation of the generative process for the Bayesian partial membership model. The plate notation represents replication of variables, and the shaded node represents observed variables. $\boldsymbol{\alpha}$ is a K-dimensional vector of positive hyperparameters. The generative model is: draw mixture weights ρ_k from a Dirichlet distribution with hyperparameters $\boldsymbol{\alpha}$, and a positive scaling factor a from an exponential distribution with hyperparameter $\beta > 0$; then draw a vector of partial memberships $\mathbf{v}_n$ from a Dirichlet distribution, representing the extent to which the observation belongs to each of the K clusters.

$$\boldsymbol{\rho} \sim \text{Dir}(\boldsymbol{\alpha}); \qquad a \sim \text{Exp}(\beta), \tag{1.7}$$

$$\mathbf{v}_n \sim \text{Dir}(a\boldsymbol{\rho}). \tag{1.8}$$

Each cluster k is characterised by an exponential family distribution with natural parameters $\boldsymbol{\theta}_k$ that are drawn from a conjugate exponential family distribution, with hyperparameters $\boldsymbol{\lambda}$ and ν. Given the latent variables and parameters, each data point is drawn from a data-appropriate exponential family distribution.

$$\boldsymbol{\theta}_k \sim \text{Conj}\left(\boldsymbol{\lambda}, \nu\right), \tag{1.9}$$

$$\mathbf{x}_n \sim \text{Expon}\left(\textstyle\sum_k v_{nk}\boldsymbol{\theta}_k\right). \tag{1.10}$$

We denote $\boldsymbol{\Omega} = \{\mathbf{V}, \boldsymbol{\Theta}, \boldsymbol{\rho}, a\}$ as the set of unknown parameters with hyperparameters $\boldsymbol{\Psi} = \{\boldsymbol{\alpha}, \beta, \boldsymbol{\lambda}, \nu\}$. Given this generative specification, the joint-

probability is:

$$p(\mathbf{X}, \mathbf{\Omega}|\mathbf{\Psi}) = p(\mathbf{X}|\mathbf{V}, \mathbf{\Theta})p(\mathbf{V}|a, \boldsymbol{\rho})p(\mathbf{\Theta}|\boldsymbol{\lambda}, \nu)p(\boldsymbol{\rho}|\boldsymbol{\alpha})p(a|\beta)$$
$$= \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{v}_n, \mathbf{\Theta})p(\mathbf{v}_n|a, \boldsymbol{\rho}) \prod_{k=1}^{K} p(\boldsymbol{\theta}_k|\boldsymbol{\lambda}, \nu)p(\boldsymbol{\rho}|\boldsymbol{\alpha})p(a|\beta). \quad (1.11)$$

Substituting the forms for each distribution, the log joint probability is:

$$\ln p(\mathbf{X}, \mathbf{\Omega}|\mathbf{\Psi}) = \sum_{n=1}^{N} \left\{ \left(\sum_k v_{nk}\boldsymbol{\theta}_k \right)^{\top} \mathbf{x}_n + h(\mathbf{x}_n) + g\left(\sum_k v_{nk}\boldsymbol{\theta}_k \right) \right\} \quad (1.12)$$
$$+ \sum_{k=1}^{K} \left[\boldsymbol{\lambda}^T \boldsymbol{\theta}_k + \nu g(\boldsymbol{\theta}_k) + f(\boldsymbol{\lambda}) \right]$$
$$+ N \ln \Gamma\left(\sum_k a\rho_k \right) - N \sum_k \ln \Gamma\left(a\rho_k\right) + \sum_n \sum_k \left(a\rho_k - 1\right) \ln v_{nk}$$
$$+ \ln \Gamma\left(\sum_k \alpha_k \right) - \sum_k \ln \Gamma(\alpha_k) + \sum_k (\alpha_k - 1) \ln \rho_k + \ln b - ba.$$

We arrived at the Bayesian Partial membership (BPM) model using a continuous latent variable relaxation of the mixture model. As a result, the BPM reduces to mixture modelling when $a \to 0$ with mixing proportions $\boldsymbol{\rho}$, and follows from the limit of eq 1.8. The BPM bears interesting relationships to several well-known models, including Latent Dirichlet Allocation (LDA) [10], mixed-membership models [15], Discrete Components Analysis [11], and exponential family PCA [13, 36], which we discuss in section 1.2.4. Unlike LDA and mixed-membership models that capture partial memberships in the form of attribute-specific mixtures, the BPM does not assume a factorisation over attributes and provides a general way of combining exponential family distributions with partial membership.

1.2.4 Exponential Family Factor Analysis

We now consider a Bayesian model for exponential family factor analysis (EXFA) [35]. We can think of an exponential family factor analysis as a method of decomposing an observed data matrix $\mathbf{X}$, which can be of any type supported by the exponential family of distributions, into two matrices $\mathbf{V}$ and $\mathbf{\Theta}$; we define the product matrix $\mathbf{P} = \mathbf{V}\mathbf{\Theta}$. Since the likelihood depends only on $\mathbf{V}$ and $\mathbf{\Theta}$ through their product $\mathbf{P}$, this can also be seen as a model for matrix factorisation. In traditional factor analysis and probabilistic PCA, the elements of the matrix $\mathbf{P}$, which are the means of Gaussian distributions, lie in the same space as that of the data $\mathbf{X}$. In the case of EXFA and similar methods for non-Gaussian PCA such as EPCA [13, 36], this matrix represents the natural parameters of the exponential family distribution of the data.

The generative process for the EXFA model is described by the graphical model of figure 1.1(b). Let $\mathbf{m}$ and $\mathbf{S}$ be hyperparameters representing a K-dimensional vector of initial mean values and an initial covariance matrix respectively. Let α and β be the hyperparameters corresponding to the shape and scale parameters of an inverse-gamma distribution. We begin by drawing $\boldsymbol{\mu}$ from a Gaussian distribution and the elements σ_k^2 of the diagonal matrix $\boldsymbol{\Sigma}$ from an inverse-gamma distribution. For each data point n of the factor score matrix $\mathbf{V}$, we draw a K-dimensional Gaussian latent variable $\mathbf{v}_n$.

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}, \mathbf{S}); \qquad \sigma_k^2 \sim i\mathcal{G}(\alpha, \beta) \tag{1.13}$$

$$\mathbf{v}_n \sim \mathcal{N}(\mathbf{v}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{1.14}$$

The data is described by an exponential family distribution with natural parameters given by the product of the latent variables $\mathbf{v}_n$ and parameters $\boldsymbol{\theta}_k$. The exponential family distribution modelling the data and the corresponding prior over the model parameters is:

$$\boldsymbol{\theta}_k \sim \text{Conj}\,(\boldsymbol{\lambda}, \nu) \tag{1.15}$$

$$\mathbf{x}_n|\mathbf{v}_n, \boldsymbol{\Theta} \sim \text{Expon}\,(\textstyle\sum_k v_{nk}\boldsymbol{\theta}_k)\,. \tag{1.16}$$

We denote $\boldsymbol{\Omega} = \{\mathbf{V}, \boldsymbol{\Theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ as the set of unknown parameters with hyperparameters $\boldsymbol{\Psi} = \{\mathbf{m}, \mathbf{S}, \alpha, \beta, \boldsymbol{\lambda}, \nu\}$. Given this specification (1.13) - (1.16), the log joint probability distribution is:

$$\begin{aligned}
p(\mathbf{X}, \boldsymbol{\Omega}|\boldsymbol{\Psi}) &= p(\mathbf{X}|\mathbf{V}, \boldsymbol{\Theta})p(\boldsymbol{\Theta}|\boldsymbol{\lambda}, \nu)p(\mathbf{V}|\boldsymbol{\mu}, \boldsymbol{\Sigma})p(\boldsymbol{\mu}|\mathbf{m}, \mathbf{S})p(\boldsymbol{\Sigma}|\alpha, \beta) \\
\ln p(\mathbf{X}, \boldsymbol{\Omega}|\boldsymbol{\Psi}) &= \sum_{n=1}^{N}\left[\left(\sum_k v_{nk}\boldsymbol{\theta}_k\right)^{\top}\mathbf{x}_n + h(\mathbf{x}_n) + g\left(\sum_k v_{nk}\boldsymbol{\theta}_k\right)\right] \\
&\quad + \sum_{k=1}^{K}\left[\boldsymbol{\lambda}^{\top}\boldsymbol{\theta}_k + \nu g(\boldsymbol{\theta}_k) + f(\boldsymbol{\lambda})\right] \\
&\quad + \sum_{n=1}^{N}\left[-\frac{K}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{v}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{v}_n - \boldsymbol{\mu})\right] \\
&\quad - \frac{K}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{S}| - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{m})^T\boldsymbol{S}^{-1}(\boldsymbol{\mu} - \boldsymbol{m}) \\
&\quad + \sum_{i=1}^{K}\left[\alpha\ln\beta - \ln\Gamma(\alpha) + (\alpha - 1)\ln\sigma_i^2 - \beta\sigma_i^2\right],
\end{aligned} \tag{1.17}$$

where the functions $h(\cdot)$, $g(\cdot)$ and $f(\cdot)$ correspond to the functions of the chosen conjugate-exponential family distribution for the data.

Whereas mixture models represent membership to a single clusters, and the BPM represents partial membership to the set of clusters, EXFA explains the

data using linear combinations of all latent classes (an all-membership). EXFA thus provides a natural way of combining different exponential family distributions and producing a shared latent embedding of the data using Gaussian latent variables.

1.3 Prior to Posterior Analysis

For both the Bayesian Partial Membership (BPM) model and exponential family factor analysis (EXFA), typical tasks include prediction, missing data imputation, dimensionality reduction, and data visualisation. To achieve this, we must infer the posterior distribution $p(\mathbf{\Omega}|\mathbf{X}, \mathbf{\Psi})$, using which we can visualise the structure of the data and compute predictive distributions. Due to the lack of conjugacy, analytic computation of the posterior is not possible. Although many approximation methods exist for computing posterior distributions, we focus on Markov chain Monte Carlo (MCMC) because it provides a simple, powerful, and often surprisingly scalable family of methods. Using MCMC involves representing the posterior distribution by a set of samples, following which we use these samples for analysis, prediction and decision making.

1.3.1 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods are a general class of sampling methods based on constructing a Markov chain with the desired posterior distribution as the equilibrium distribution of the Markov chain. MCMC methods are popular in machine learning and Bayesian statistics and include widely-known methods such as Gibbs sampling, Metropolis-Hastings and slice sampling [41, 19]. For sampling in the models of section 1.2.3 and 1.2.4, we make use of a general purpose MCMC algorithm known as Hybrid (or Hamiltonian) Monte Carlo (HMC) sampling.

Hybrid Monte Carlo (HMC), which was first described by Duane et al. [14], is based on the simulation of Hamiltonian dynamics as a way of exploring the sample space of the posterior distribution. Consider the task of generating samples from the distribution $p(\mathbf{\Omega}|\mathbf{\Psi}, \mathbf{X})$, with $\mathbf{\Psi}$ being any relevant hyperparameters; we denote $\mathbf{u}$ as an auxiliary variable. Intuitively, HMC combines auxiliary variables with gradient information from the joint-probability to improve mixing of the Markov chain, with the gradient acting as a force that results in more effective exploration of the sample space. HMC can be used to sample from continuous distributions for which the density function can be evaluated (up to a known constant). This makes HMC particularly amenable to sampling in non-conjugate settings where the full conditional distributions

required for Gibbs sampling cannot be derived, but for which the joint probability density and its derivatives can be computed. These properties make HMC well suited to sampling from the BPM and EXFA models, since these models do not have a conjugate structure and all unknown variables $\mathbf{\Omega}$ are continuous and differentiable, making it possible to exploit available gradient information.

For HMC, a potential energy function and a kinetic energy function is defined, whose sum forms the Hamiltonian energy:

$$\mathcal{H}(\mathbf{\Omega}, \mathbf{u}) = \mathcal{E}(\mathbf{\Omega}|\mathbf{\Psi}) + \mathcal{K}(\mathbf{u}), \qquad \text{(Hamiltonian Energy)} \qquad (1.18)$$

$$\mathcal{E}(\mathbf{\Omega}|\mathbf{\Psi}) = -\ln p(\mathbf{\Omega}, \mathbf{X}|\mathbf{\Psi}), \qquad \text{(Potential Energy)} \qquad (1.19)$$

$$\mathcal{K}(\mathbf{u}) = -\tfrac{1}{2}\mathbf{u}^\top \mathbf{M}\mathbf{u}. \qquad \text{(Kinetic Energy)} \qquad (1.20)$$

The Hamiltonian can be seen as the log of an augmented distribution to be sampled from: $p(\mathbf{X}, \mathbf{\Omega}, \mathbf{u}|\mathbf{\Psi}) = p(\mathbf{X}, \mathbf{\Omega}|\mathbf{\Psi})\mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{M})$, where $\mathbf{M}$ is a preconditioning matrix often referred to as a mass matrix, which in the simplest case is set to the identity matrix. The gradient of the potential energy is defined as: $\Delta(\mathbf{\Omega}) = \frac{\partial \mathcal{E}(\mathbf{\Omega})}{\partial \mathbf{\Omega}}$. We defer further details of the physical underpinnings describing Hamiltonian dynamics and its appropriateness for MCMC to the work of MacKay [38] and Neal [37].

We present the full algorithm for HMC in algorithm 1.1. Each iteration of HMC has two steps. In the first step, we assume that an initial sample (state) for $\mathbf{\Omega}$ is given and generate a Gaussian variable $\mathbf{u}$ for the momentum (line 4, algorithm 1.1). In the second step, we simulate Hamiltonian dynamics, which follows the equations of motion to move the current sample and momentum to a new state. The Hamiltonian dynamics must be discretised for implementation and the most popular discretisation is known as the leapfrog method (lines 7-11). The leapfrog approximation is simulated for L steps using a step-size ϵ. The samples $\mathbf{\Omega}^*$ and $\mathbf{u}^*$ at the end of the leapfrog steps form the proposed state, which is accepted using the Metropolis criterion (line 15):

$$\min\left(1, \exp(-\mathcal{H}(\mathbf{\Omega}^*, \mathbf{u}^*) + \mathcal{H}(\mathbf{\Omega}, \mathbf{u}))\right). \qquad (1.21)$$

Finally, marginal samples from $p(\mathbf{\Omega})$ are obtained by ignoring $\mathbf{u}$.

1.3.1.1 Aspects of Implementation

To implement HMC correctly we must adjust the energy function to account for variables that may be constrained, such as variables that are non-negative or bound between [0,1]. We make use of the following transformations:

BPM with $a > 0$, $\sum_k \pi_k = 1$ and $\sum_k v_{nk} = 1$:

$$a = \exp(\eta); \qquad \pi_k = \frac{\exp(r_k)}{\sum_{k'} \exp(r_{k'})}; \qquad v_{nk} = \frac{\exp(\omega_{nk})}{\sum_{k'} \exp(\omega_{nk'})}.$$

Algorithm 1.1: Hybrid Monte Carlo (HMC) Sampling [31]

Evaluate Gradient $\mathbf{g} = \boldsymbol{\Delta}(\boldsymbol{\theta})$ with initial $\boldsymbol{\theta}$ `// g = gradE(theta)`
Evaluate Energy $E = \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\psi})$ `// E = findE(theta)`
for L *iterations* **do**
  Initialise new momentum $\mathbf{u}$ drawn from a Gaussian
  Calculate: $\mathcal{K}(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top\mathbf{u}$ and $\mathcal{H} = \mathcal{E}(\boldsymbol{\theta}|\boldsymbol{\psi}) + \mathcal{K}(\mathbf{u})$
  $\boldsymbol{\theta}^{new} \leftarrow \boldsymbol{\theta}$; $\boldsymbol{g}^{new} \leftarrow \boldsymbol{g}$;
  for L *leapfrog steps* **do**
    $\mathbf{u} \leftarrow \mathbf{u} - \frac{\epsilon}{2}\boldsymbol{g}$ `// Make half-step in` u
    $\boldsymbol{\theta}^{new} \leftarrow \boldsymbol{\theta}^{new} + \epsilon\mathbf{u}$ `// Make a step in theta`
    $\boldsymbol{g}^{new} \leftarrow \boldsymbol{\Delta}(\boldsymbol{\theta}^{new})$ `// gradE(thetaNew)`
    $\mathbf{u} \leftarrow \mathbf{u} - \frac{\epsilon}{2}\boldsymbol{g}^{new}$ `// make half step in` u
  $E^{new} = \mathcal{E}(\boldsymbol{\theta}^{new}|\boldsymbol{\psi})$ `// Enew = findE(thetaNew)`
  Calculate $\mathcal{K}(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top\mathbf{u}$
  Hamiltonian $\mathcal{H}^{new} \leftarrow E^{new} + \mathcal{K}(\mathbf{u})$
  if *rand()* $< \exp(-(\mathcal{H}^{new} - \mathcal{H}))$ **then**
    Accept $\leftarrow$ `True`
    $\boldsymbol{g} \leftarrow \boldsymbol{g}^{new}$; $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{new}$; $E \leftarrow E^{new}$
  else
    Accept $\leftarrow$ `False`

EXFA with $\sigma_k^2 > 0$: $\sigma_k^2 = \exp(\xi_k)$.

The use of these transformations requires the inclusion of the determinant of the Jacobian of the change of variables, as well as consistent application of the chain rule for differentiation taking into account the change of variables.

HMC has two tunable parameters: the number of leapfrog steps L and the step-size ϵ. In general the step-size should be chosen to ensure that the sampler's rejection rate is between $25\% - 35\%$, and to use a large number of leapfrog steps. We generally make use of L between 80 and 100 here. The tuning of these parameters can be challenging in some cases, and we show ways in which these choices can be explored in the experimental section. We fix the mass matrix to the identity but this can also be tuned, and we discuss aspects of this in section 1.6. Analysis of the optimal acceptance rates for HMC is discussed by Beskos et al. [4]; the work of Neal [38] provides a great deal of guidance in tuning HMC samplers.

Many data sets contain missing values, and we can account for this missing data in a principled manner by dividing the data into the set of observed and missing entries $\mathbf{X} = \{\mathbf{X}^{obs}, \mathbf{X}^{missing}\}$, and conditioning on the set $\mathbf{X}^{obs}$ during inference. In practice, the pattern of missing data is represented by a

masking matrix, which is the indicator matrix of elements that are observed versus missing. Probabilities are then computed using elements of the masking matrix set to one.

1.4 Related Work

Mixed-membership models and LDA: In general, mixed-membership models [15] organise the data in two levels using an admixture structure (mixture-of-mixtures model). Latent Dirichlet Allocation (LDA) [10], as an instance of a mixed-membership model, organises the data at the level of words and then documents, expressing this data likelihood as a mixture of multinomials. LDA combines this mixture-likelihood with a K-dimensional Dirchlet-distributed latent variable $\mathbf{v}$ as a distribution over topics. The BPM is a similar *latent Dirichlet model*, but the latent variable represents partial memberships, and instead of a two-level structure, the BPM indexes the data directly using an exponential family likelihood. LDA assumes that each data attribute (i.e. words) of an observation (i.e. document) is drawn independently from a mixture distribution given the membership vector for the data point, $x_{nd} \sim \sum_k v_{nk} p(x|\theta_{kd})$. As a result, LDA makes most sense when the observations (documents) being modelled constitute bags of exchangeable sub-objects (words). Furthermore, for both LDA and mixed-membership models, there is a discrete latent variable for every sub-object, corresponding to which mixture component that sub-object was drawn from. This large number of discrete latent variables makes MCMC sampling potentially much more expensive than sampling in the exponential family models we describe here. A more detailed discussion and comparison of mixed- and partial-membership models in the chapter by Gruhl and Erosheva [22, §1.4] in this volume complements this discussion.

Latent Gaussian models: EXFA employs a K-dimensional Gaussian latent variable $\mathbf{v}$, and is thus an example of a *latent Gaussian model*. This is one of the most established classes of models, and includes generalised linear regression models, non-parametric regression using Gaussian processes, state-space and dynamical systems, unsupervised latent variable models such as PCA, Factor Analysis [3] and Probabilistic Matrix Factorisation [43], and Gaussian Markov random fields. In generalised linear regression [7], the latent variables v_n are the predictors formed by the product of covariates and regression coefficients; in Gaussian process regression [40], the latent variables $\mathbf{v}$ are drawn jointly from a correlated Gaussian using a mean function and a covariance function formed using the covariates; and in probabilistic PCA and factor analysis [47, 3], latent variables $\mathbf{v}_n$ are Gaussian with isotropic or diagonal covariances, respectively.

EXFA also follows as a Bayesian interpretation of exponential family PCA [13] and generalised latent trait models [36]. Instead of fully Bayesian inference, these related models specify an objective function that is optimised to obtain the MAP solution. Similarly to the BPM, in EXFA, the data is indexed directly using an exponential family distribution, rather than through an admixture structure. With this realisation though, it is easy to see the connection and extension of EXFA to a generalised mixed-membership matrix factorisation (MMMF) models, by instead considering a two level representation of the data similar to that described by Mackey at al. [32].

Both the BPM and EXFA model the natural parameters of an exponential family distribution. This makes them different from other latent variable models, such as non-negative matrix factorisation (NMF) [29, 11], since these alternative approaches model the mean parameters of distributions rather than their natural parameters. The use of natural parameters allows for easier learning of model parameters, since these are often unconstrained, unlike learning for NMF, which requires special care in handling constraints, e.g., leading to the multiplicative updates required for learning in NMF.

Fuzzy clustering: Partial membership is a cornerstone of fuzzy theory, and the notion that probabilistic models are unable to handle partial membership is used to argue that probability is a sub-theory, or different in character from fuzzy logic [49, 28]. With the BPM, we are able to demonstrate that probabilistic models *can* be used to describe partial membership. Rather than using a mixture model for clustering, an alternative is given by fuzzy set theory and fuzzy k-means clustering [5]. Fuzzy k-means clustering [18] iteratively minimises the objective function: $J = \sum_n \sum_k v_{nk}^{\gamma_f} D^2(\mathbf{x}_n, \mathbf{c}_k)$, where $\gamma_f > 1$ is the fuzzy exponent parameter, v_{nk} represents the degree of membership of data point n to cluster k, where $\sum_k v_{nk} = 1$ and $D^2(\mathbf{x}_n, \mathbf{c}_k)$ is a squared distance between the observation $\mathbf{x}_n$ and the cluster centre $\mathbf{c}_k$. By varying γ_f it is possible to attain different degrees of partial membership, with $\gamma_f = 1$ being k-means with no partial membership.

We compare fuzzy clustering and the BPM in section 1.5 and find that the two approaches achieve very similar results, with the advantage of probabilistic models being that we obtain estimates of uncertainty, are able to deal with missing data, and can combine these models naturally with the wider set of probabilistic models. Thus, we hope that this work demonstrates that, contrary to the common misconception, fuzzy set theory is not needed to represent partial membership in probabilistic models, and that this can be achieved with established approaches for probabilistic modelling.

1.5 Experimental Results

We demonstrate the effectiveness of the models presented in this chapter using synthetic data sets as well with a real world case study roll call data from the US senate. We evaluate the performance of the methods by computing the negative log predictive probability (NLP) on test data. The test sets are created by setting 10% of the elements of the data matrix as missing data in the training set, and then learning in the presence of this missing data. We provide Matlab code to reproduce all the results in this section online[1].

1.5.1 Synthetic Binary Data

1.5.1.1 Noisy Bit Patterns

We evaluate the behaviour of EXFA using a synthetic binary data set. The synthetic data was generated by creating three 16-bit prototype vectors, with each bit being set with probability 0.5. Each of the three prototypes is replicated 100 times, resulting in a data set of 300 observations. Noise is then added to the data by flipping bits in the data set with probability of 0.1 [46, 35]. We use HMC to generate 5000 samples from the EXFA model with $K = 3$ factors, and demonstrate the evolution of the sampler in figure 1.2. Since the sampler is initialised randomly, we see that the initial samples have no discernible structure. As the sampling proceeds, the energy rapidly decreases (the energy is the negative log joint probability, meaning lower is better), and useful structure can be seen after the 500th sample. By the end of the sampling, the samples correctly capture the true data, as seen by comparing the mean reconstruction computed using the last 1000 samples, and the true data in figure 1.2. The predictive probability of the test data computed for every sample also decreases as the sampler progresses, indicating that the correct latent structure has been inferred, allowing for accurate imputation of the missing data. The random predictor would have an $NLP = 10\% \times 300 \times 16 = 480$ bits, and we can see that the NLP we obtain is much lower than this. The maximum likelihood estimation of EXFA has $NLP = 1148$ bits, which is significantly worse than the Bayesian prediction. This is a well known problem, since maximum likelihood estimation in this model suffers from severe overfitting, highlighting an important advantage of Bayesian methods over optimisation methods [35].

Plots such as figure 1.2 are also useful as tools for tuning an HMC sampler. For a fixed K, the region of high energy is fixed, so this can be used to choose a step-size and number of leapfrog steps that allows us to rapidly reach this region. We fix $L = 80$ and tune ϵ by monitoring the progression of the sampler.

[1] `www.shakirm.com/code/EFLVM/`

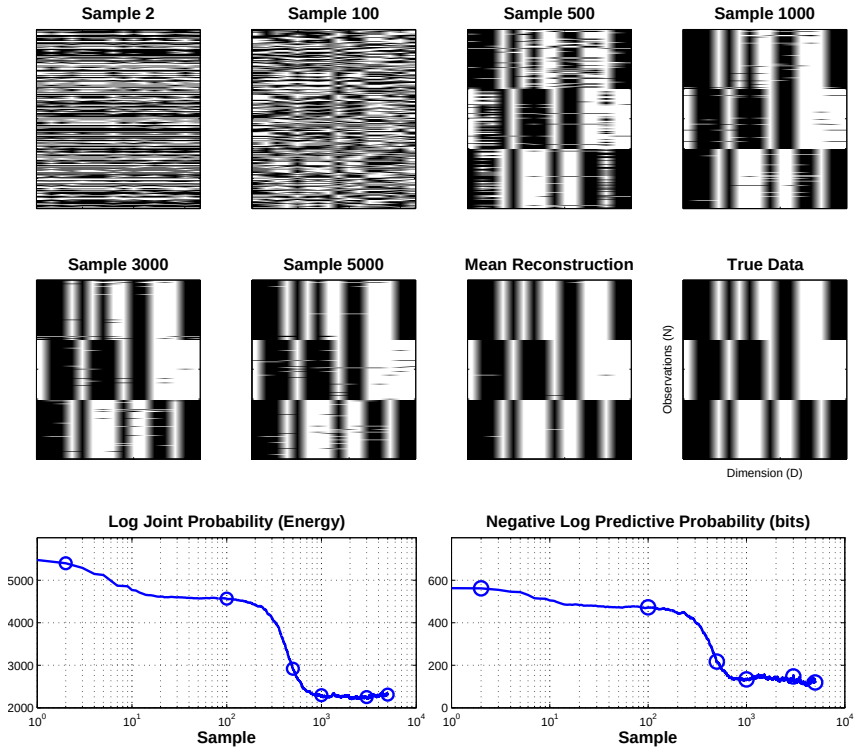


FIGURE 1.2
Reconstruction of data samples at various stages of the sampling in EXFA. Top two rows: Greyscale reconstructions at various samples and the true, noise-free data. Bottom row: Change in the energy function (using training data) and the corresponding predictive probability (using test data). We show circular markers at samples for which the reconstructions are shown above.

In practice, we can choose the number of latent factors K by cross-validation. To do this, we create 10 replications of our data and for each data set we set 10% of the elements of the matrix as missing, using these elements as a held-out data set. We then generate samples from the model over a range of K, and use the reconstruction error on the held-out data to choose the K that gives the best performance. We compare the negative log predictive probability (NLP) for K in the range of 2 to 20. We show the performance on the training and testing data in terms of root mean squared error (RMSE) as well as predictive probability (NLP) in figure 1.3. We also compare the performance of the fully Bayesian approach using HMC that we presented, and the performance of maximum likelihood estimation in this model. The maximum likelihood estimators experience severe overfitting as shown by the RMSE on the training data. Since we would prefer a simpler model to a more

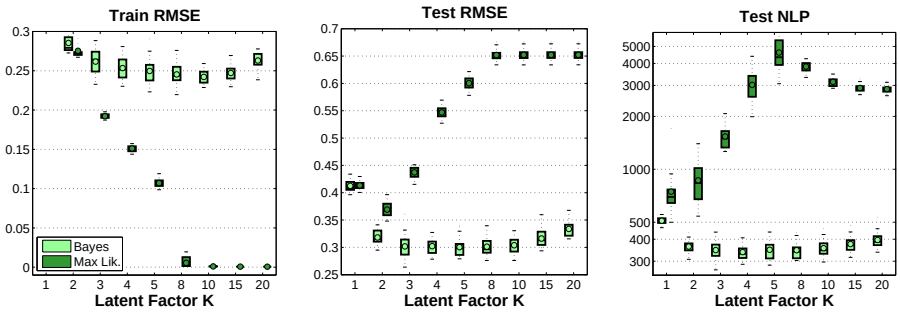


FIGURE 1.3
Choosing the number of latent factors K by cross-validation. We find that $K = 3$ is an appropriate number of latent factors.

complex one, we choose $K = 3$ vased on the graphs of RMSE and NLP on the test data. We discuss this issue of selecting K, and in particular, automatic methods for its selection in section 1.6.

1.5.1.2 Simulated Data from the BPM

We generated a synthetic binary data set from the BPM , consisting of $N = 50$ points, each being a $D = 32$ dimensional vector, using $K = 3$ clusters. We ran HMC for 4000 iterations, using the first half as burnin. To compare the true partial memberships $\mathbf{V}_T$ to the inferred memberships $\mathbf{V}_L$, we compute $\mathbf{U}_T = \mathbf{V}_T\mathbf{V}_T^\top$ and $\mathbf{U}_L = \mathbf{V}_L\mathbf{V}_L^\top$, which is a measure of the degree of shared membership between pairs of observations for the true and inferred partial memberships, respectively [23]. This measure is invariant to permutations of the cluster labels, and the range of entries is between [0,1]. We show image-maps of these matrices in figure 1.4. The difference between entries of the true and inferred shared memberships $|\mathbf{U}_T - \mathbf{U}_L|$ is shown in the histogram. The two matrices are highly similar, with 90% of entries being different from the true value by less than 0.2, showing that the sampler was able to learn the true partial memberships.

1.5.2 Senate Roll Call Data

Having evaluated the behaviour of the BPM and EXFA on synthetic data, we demonstrate their use in exploring membership behaviour from the US Senate roll call as a case study. Specifically, we analyse the roll call from the 107th US congress (2001-2002) [25]. The data consists of 99 senators (one senator died in 2002, and neither he nor his replacement are included), by 633 votes. It also includes the outcome of each vote, which we treat as an additional data points (like a senator who always voted the actual outcome). The matrix contains binary features for yea and nay votes, and abstentions are recorded as missing

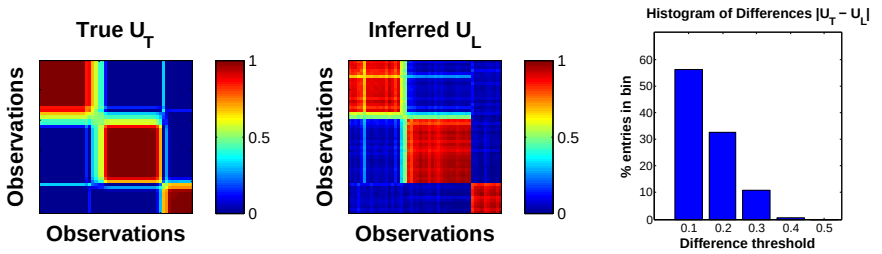


FIGURE 1.4
Image-maps showing true shared partial memberships $\mathbf{U}_T$ and inferred shared membership $\mathbf{U}_L$ for synthetic data generated from the BPM model. The histogram shows the percentage of entries in $|\mathbf{U}_T - \mathbf{U}_L|$ that fall within a given difference threshold.

values. For the perspective of a political scientist on analysisng such data, see the chapter by Gross and Manrique-Vallier [21].

We analyse the data using the BPM with $K = 2$ clusters, and show results of this analysis in figure 1.5. Since there are two clusters and the amount of membership always sums to 1 across clusters, the figure looks the same regardless of whether we look at the 'Democrat' or 'Republican' cluster. The cyan line in figure 1.5 indicates the partial membership assigned to each of the senators with their names overlaid. We can see that most Republicans and Democrats are clustered together in the flat regions of the line (with partial memberships very close to 0 or 1), but that there is a fraction of senators (around 20%) that lie somewhere in-between. Interesting properties of this figure include the location of Senator Jeffords (in magenta) who left the Republican party in 2001 to become an Independent who caucused with the Democrats. Also Senator Chafee who is known as a moderate Republican and who often voted with the Democrats (for example, he was the only Republican to vote against authorising the use of force in Iraq), and Senator Miller a conservative Democrat who supported George Bush over John Kerry in the 2004 US Presidential elections. Lastly, it is interesting to note the location of the Outcome data point, which is very much in the middle. This makes sense since the 107th congress was split 50-50 (with Republican Dick Cheney breaking ties), until Senator Jeffords became an Independent at which point the Democrats had a one seat majority.

We also analysed the data using fuzzy k-means clustering, which found very similar rankings of senators to the 'Democrat' cluster. Fuzzy k-means was very sensitive to the exact ranking and degree of partial membership, since it is highly sensitive to the fuzzy exponent parameter γ_f, which is typically set by hand. Figure 1.6 shows the change in partial membership for the outcome and the most-allegiant Democrat and Republican senator (using the result

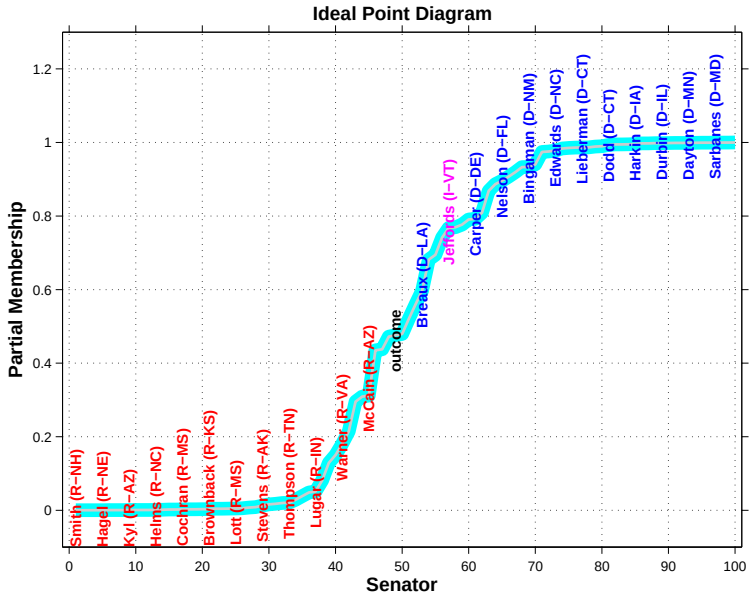


FIGURE 1.5
Analysis of the partial memberships for the 107th US Senate roll call using BPM. The line shows the amount of membership in the 'Democrat' cluster with the names of Democrat senators overlaid in blue and Republican senators in red.

of figure 1.5), for a range of values for the fuzzy exponent. The graph shows that the assigned partial membership can vary quite dramatically depending on the choice of γ_f. This type of sensitivity to parameters does not exist in the Bayesian models we present here, since they can be inferred automatically.

The BPM provides a very natural representation of the membership of individuals in this data to political leanings. An alternative viewpoint can be obtained using EXFA. With EXFA, the latent variables do not have an interpretation as a degree of membership, but rather provide a low-dimensional embedding of the data, which for the case of 2 latent factors, can be used to provide a spatial visualisation of senators. We show the results of analysing the roll call data with EXFA in figure 1.8, using $K = 2$ latent factors, producing 4000 samples from the HMC sampler, using the first half as burnin. The latent embedding in figure 1.8 is colour-coded blue for Democrats and red for Republicans, and shows that there is a natural separation of the data into these two groups. Similarly to the BPM, we observe that most senators are clustered into a Democrat or Republican cluster, with a percentage who straddle the boundary between these two groups. Again, we see the effect of

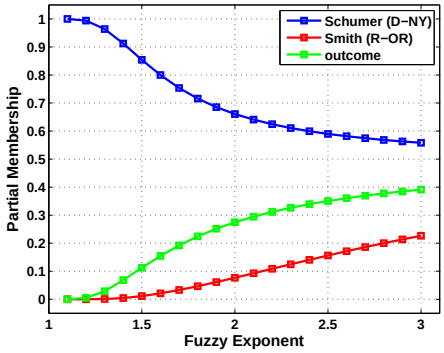


FIGURE 1.6
Sensitivity of partial memberships in fuzzy k-means with respect to the fuzzy exponent.

NLP	BPM	EFA	DPM
Mean	192	188	196
Min	100	92	112
Median	173	171	178
Max	428	412	412
Outcome	230	183	245

FIGURE 1.7
Comparison of negative log predictive probabilities (in bits) across senators for BPM, EFA and DPM.

the independent candidate and the outcome. It is also important to note the connection between both BPM and EFA to ideal point models in political science [2], which aim to spatially represent political preferences on a left-to-right scale. Using the BPM and EFA, we have with figures 1.5 and 1.8 shown Bayesian approaches of producing 1D and 2D ideal point representations, respectively.

As a further comparison of the BPM and EXFA, we also analyse the roll call data using a Dirichlet Process mixture model (DPM). We ran the DPM for 1000 Gibbs sampling iterations, sampling both assignments and concentration parameters. The DPM confidently finds 4 clusters: one cluster consists soley of Democrats, another solely of Republicans, a third cluster contains 9 moderate Republicans and Democrats as well as the outcome, and the last cluster consists of a single senator (Hollings (D-SC)).

We calculate the negative log-predictive probability (NLP, in bits) across senators for the BPM, EXFA and DPM (Table 1.7). We present the mean, minimum, median and maximum NLP over all senators, which represents the number of bits needed to encode a senator's voting behaviour. We also show the outcome separately. Except for the maximum, the BPM is able to produce a more compressed representation for each senator than the DPM, showing the sensibility of inferring partial memberships for this data, rather than assignments to clusters. EXFA produces the most compressed representation, since it used unconstrained latent variables and thus has greater modelling flexibility. These two approaches emphasise the tradeoff between modelling efficiency and interpretability that must be considered when analysing such data.

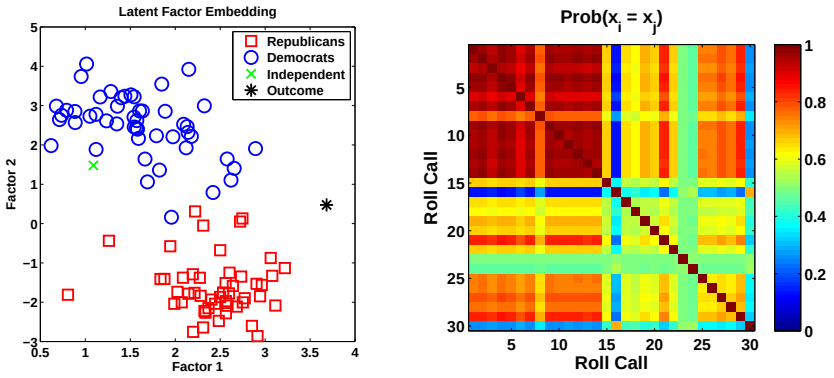


FIGURE 1.8
Analysis of the partial memberships for the 107th US Senate roll call using EXFA. (a) The left plot shows the latent embedding produced using 2 latent factors. (b) The right plot shows the marginal covariance between votes.

The BPM gave an intuitive numerical quantity to the degree of membership, whereas EXFA gives an intuitive spatial understanding of this membership. Factor models are also often used to model the covariance structure of data and provide further insight into the data. For Gaussian data, this covariance is given by $\mathbf{\Theta}\mathbf{\Theta}^\top$. For non-Gaussian data, we can compute the marginal covariance $p(x_i = x_j), i \neq j$, by Monte Carlo integration using the posterior samples obtained. We show this in figure 1.8(b) for the first 30 votes. The figure shows that there are many roll calls that are highly correlated, e.g., the first 14 entries represent the opening of the congress and are votes for chairs of various committees. Often not being votes of contention, there is highly correlated voting for these motions. Analysis of this matrix gives insight into the evolution of votes in the congress and provides an example of some of the probabilistic queries that can be made once the posterior samples are obtained. Other interesting probabilistic queries of this nature include examining the similarity of senators using the KL-distance between their latent posterior distributions, or examining the influence of senators to the voting outcomes using the marginal likelihood each senator contributes to the total probability.

1.6 Discussion

Having gained an understanding of exponential family latent variable models and their behaviour, we now consider some of the questions that affect

our ability to use such models in practice. Question that arise include: how to decide between competing models; methods for choosing the latent dimensionality K; difficulty in tuning the MCMC samplers; and obstacles in applying these models to large data sets. We expand on these questions and discuss the ways in which our models can be extended to address them.

Choice of model. In this chapter we have considered mixture models, the Bayesian partial-membership model, exponential family factor analysis, and mixed-membership models. The choice of one model type over another depends on the whether the modelling assumptions made match our beliefs regarding the process that generated the data, as well the aim of our modelling effort, whether for visualisation, predictive, or explanatory purposes. The BPM and EXFA are models with a single layer of latent variables that we showed are relaxations of K-component mixture models. These models thus make use of a single layer of latent variables, and we demonstrated in the experiments that the models allowed for de-noising of data, effective imputation of missing data, and are useful tools for visualisation of high-dimensional data. The structure of the models proved to be intuitive and flexible, and appropriate for the tasks we presented.

More flexible versions of these models can be obtained by considering the mixed-membership analogues of the BPM and EXFA, such as grade-of-membership models [16, 21] and mixed-membership matrix factorisation [32], respectively. In addition, other prior assumptions may be needed; sparsity is one such prior assumption that has gained importance and the inclusion of sparsity in the models discussed here is described by Mohamed et al. [34]. The chapter by Galyardt [17] in this volume shows that mixed-membership models have an equivalent representation as a mixture model, with a number of mixture components polynomial in K, thus providing a highly efficient representation of high dimensional data. Inference in these more complex models is harder due to the increased number of latent and assignment variables, making the factors affecting our choice of model based on the tradeoff between simplicity, flexibility and the computational complexity of the available models. A formal model comparison would rely on Bayesian model selection, in which the 'best' model is chosen based on the evaluation of the marginal likelihood or model evidence [12].

Choosing the latent dimensionality. In section 1.5 we used cross-validation to determine the appropriate dimensionality of the latent variables. Ideally, we would wish to learn K automatically using the training data only. An alternative approach to cross-validation is by Bayesian model selection where we evaluate and compare the marginal likelihood or evidence for various models, e.g., as described by Minka [33] for probabilistic PCA. The models we have described can also be adapted to include the determination of K as part of the learning algorithm. Bishop [8] exploited sparsity by employing

Automatic Relevance Determination (ARD), which uses are large number of latent factors and sets to zero any factors that are not supported by the data; K is then the number of non-zero columns at convergence of the algorithm. It is also possible to specify the dimensionality of the latent variables as part of our model construction. This approach requires an efficient means of sampling in spaces with changing dimensionality, most often achieved by trans-dimensional MCMC, such as the approach described by Lopes and West [30]. More recent approaches have focused on the construction of non-parametric latent factor models using the Indian Buffet Process or other non-parametric priors to automatically adapt the dimensionality of latent variables [27, 6].

Tuning MCMC samplers. We made use of the standard approach for Hybrid Monte Carlo (HMC) sampling here, but this can be improved to increase the number of uncorrelated samples obtained. We used of an identity mass matrix, but adaptively estimating the mass matrix, using the empirical covariance or Hessian of the log-joint probability from the samples during the burn-in phase can be used, reducing sensitivity to the choice of step-size ϵ [1]. Using an appropriate mass matrix allows proposals to be made at an appropriate scale, thus allowing for larger step-sizes during sampling. But estimation of the mass matrix (and computing its inverse) can add significantly to the computation involved in HMC. Adaptive tuning of the mass matrix was also shown using the Riemann geometry of the joint-probability by Girolami and Calderhead [20]. Another way of improving HMC was proposed by Shahbaba et al. [44], and involves splitting the Hamiltonian in a way that allows much of the movement around the state-space to be done at low computational cost [44]. Tuning the HMC parameters can be challenging, especially for the non-expert, and methods now exist for the automatic tuning of HMC's parameters [24, 48]. Any of these approaches removes the need for tuning HMC and have the promise of making the application of HMC much more general purpose.

Deterministic approximations for large-scale learning. With the increasing size of data sets, the availability of scalable inference is an important factor in the practial use of many models. MCMC methods can be shown to scale well to large data sets [43]. Deterministic approximations are increasingly used in the development of scalable algorithms, and can allow better exploitation of the distributed nature of modern computing environments. Variational inference for LDA was described by Teh et al. [45], and such an approach can be applied to the BPM. For latent Gaussian models, approximate inference methods such as Integrated Nested Laplace Approximations (INLA) [42] have been proposed. INLA is effective for models whose latent variables are controlled by a small number of hyperparameters, limiting the application of this approach for learning in EFA. Variational methods for EXFA have also been successfully explored [26].

1.7 Conclusion

In this chapter, we have described a principled Bayesian framework for latent variable modelling that is generalised to the exponential family of distributions. We began with the widely-used mixture model and showed that a relaxation of the assumption that each data point belongs to one and only one cluster allows us to explore different aspects of the structure underlying the data. We obtained the Bayesian Partial Membership (BPM) model by allowing the latent variables to represent fractional membership in multiple clusters, and obtained exponential family factor analysis (EXFA) by considering continuous latent variables, which explain contributions to the data using a linear combination from all clusters. By framing these models in the same latent variable framework, we exploited the continuous nature of the unknown parameters and demonstrated how Hybrid Monte Carlo can be implemented and tuned for such models. We also described the connection to other latent variable and mixed-membership models. Using both synthetic and real-world data, we demonstrated the use of these models for visualisation and predictive tasks and the wide range of insightful probabilistic queries that can be made using these models.

Bibliography

[1] Y. Atchade, G. Fort, E. Moulines, and P. Priouret. *Bayesian Time Series Models*, chapter Adaptive Markov Chain Monte Carlo: Theory and Methods. Cambridge University Press, 2011.

[2] J. Bafumi, A. Gelman, D. K. Park, and N. Kaplan. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13(2):171–187, 2005.

[3] D. J. Bartholomew and M. Knott. *Latent variable models and factor analysis*, volume 7 of Kendall's library of statistics. Arnold, 2nd edition, 1999.

[4] A. Beskos, N. Pillai, G. O. Roberts, J.-M. Sanz-Serna, and A. M. Stuart. Optimal tuning of hybrid Monte Carlo. http://arxiv.org/abs/1001.4460, 2010.

[5] J. Bezdek. *Pattern Recognition with Fuzzy Objective Functions algorithms*. Kluwer, 1981.

[6] A. Bhattacharya and D. B. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011.

[7] P. J. Bickel and K. A. Doksum. *Mathematical statistics: Basic ideas and selected topics*, volume I. Prentice Hall, 2001.

[8] C. M. Bishop. Bayesian PCA. In *Neural Information Processing Systems (NIPS)*, pages 382–388, 1999.

[9] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, August 2006.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993 – 1022, 2003.

[11] W. Buntine and A. Jakulin. Discrete components analysis. In *Subspace, Latent Structure and Feature Selection*, volume 3940/2006, pages 1–33. Springer (LNCS), 2006.

[12] B. P. Carlin and S. Chib. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of Royal Statistical Society Series B*, pages 473–484, 1995.

[13] M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal component analysis to the exponential family. In *Advances in Neural Information Proceeding Systems*, pages 617–624, 2002.

[14] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195:216 – 222, 1987.

[15] E. Erosheva, S. Feinberg, and J. Lafferty. Mixed membership models of scientific publications. *Proceedings of the National Academy of Science, USA*, 2004.

[16] E. A Erosheva, S. E. Fienberg, and C. Joutard. Describing disability through individual-level mixture models for multivariate binary data. *The annals of applied statistics*, 1(2):346, 2007.

[17] A. Galyardt. *Handbook of Mixed-membership models*, chapter Interpreting mixed membership models: Implications of Erosheva's representation theorem. Chapman & Hall / CRC Press, 2013.

[18] A. Gasch and M. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3, 2002.

[19] W. R. Gilks, S. Richardson, and D. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics (Chapman & Hall/CRC Interdisciplinary Statistics)*. Chapman and Hall/CRC, 1995.

[20] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of Royal Statistical Society Series B*, 73(2):123–214, 2011.

[21] J. H. Gross and D. Manrique-Vallier. *Handbook of Mixed-membership models*, chapter A mixed-membership approach to the assessment of political ideology from survey responses. Chapman & Hall / CRC Press, 2013.

[22] J. Gruhl and E. Erosheva. *Handbook of Mixed-membership models*, chapter A Tale of two (types of) mixed memberships: Comparing mixed and partial membership with a continuous data example. Chapman & Hall / CRC Press, 2013.

[23] K. A. Heller, S. Williamson, and Z. Ghahramani. Statistical models for partial membership. In *Proceedings of the International Conference on Machine Learning*, pages 392–399, 2008.

[24] M. D. Hoffman and A. Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. Technical report, http://arxiv.org/abs/1111.4246, 2011.

[25] A. Jakulin. `stat.columbia.edu/∼jakulin/Politics/`. 2002.

[26] M. E. Khan, B. Marlin, G. Bouchard, and K. Murphy. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Proceeding Systems*, 2010.

[27] D. Knowles and Z. Ghahramani. Nonparametric Bayesian sparse factor models with application to gene expression modelling. *Annals of Applied Statistics*, In Press, 2010.

[28] B. Kosko. *Neural networks and fuzzy systems*. Prentice Hall, 1992.

[29] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(21):788 –791, 1999.

[30] H. F. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–68, 2004.

[31] D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, June 2003.

[32] L. Mackey, D. Weiss, and M. I. Jordan. Mixed membership matrix factorisation. In *Proceedings of the International Conference on Machine Learning*, 2010.

[33] T. P Minka. Automatic choice of dimensionality for PCA. In *Advances in Neural Information Proceeding Systems*, pages 598–604. MIT; 1998, 2001.

[34] S. Mohamed, K. A. Heller, and Z. Ghahramani. Bayesian and L1 approaches for sparse unsupervised learning. In *Proceedings of the International Conference on Machine Learning*, 2012.

[35] S. Mohamed, K.A. Heller, and Z. Ghahramani. Bayesian Exponential Family PCA. In *Advances in Neural Information Proceeding Systems*, 2008.

[36] I. Moustaki and M. Knott. Generalized latent trait models. *Psychometrika*, 65(3):391–411, 2000.

[37] R. M. Neal. Probabilistic inference using Markov Chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto, 1993.

[38] R. M. Neal. *Handbook of Markov Chain Monte Carlo*, chapter "MCMC using Hamiltonian dynamics". Chapman & Hall / CRC Press, 2010.

[39] S. Newcomb. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8:343–366, 1886.

[40] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[41] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2nd edition, 2004.

[42] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of Royal Statistical Society Series B*, 71(2):319–392, 2009.

[43] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the International Conference on Machine Learning*, volume 25, 2008.

[44] B. Shahbaba, S. Lan, W. O. Johnson, and R. M. Neal. Split Hamiltonian Monte Carlo. Technical report, University of California, Irvine. arXiv:1106.5941, 2011.

[45] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Proceeding Systems*, volume 19, page 1353. MIT, 2007.

[46] M. E. Tipping. Probabilistic visualisation of high dimensional binary data. In *Advances in Neural Information Proceeding Systems*, volume 11, pages 592 – 598, 1999.

[47] M. E. Tipping and C. M. Bishop. Probabilistic principal components analysis. Technical Report NCRG/97/010, Neural Computing Research Group, Aston University, 1997.

[48] Z. Wang, S. Mohamed, and N. de Freitas. Adaptive Hamiltonian-based Monte Carlo samplers. In *International Conference on Machine Learning*, 2013.

[49] L. Zadeh. Fuzzy sets. *Information and Control*, 8, 1965.