A Simple and General Exponential Family Framework for Partial Membership and Factor Analysis

Zoubin Ghahramani
Department of Engineering, University of Cambridge, Cambridge CB2 1TN, United Kingdom

Shakir Mohamed
Google London, London SW1W 9TQ, United Kingdom

Katherine Heller
Department of Statistical Science, Duke University, Durham, NC 27708, USA

CONTENTS

4.1 Introduction ................................................................................ 68
4.2 Membership Models for the Exponential Family ................................. 69
  4.2.1 The Exponential Family of Distributions .................................... 69
  4.2.2 Beyond Mixture Models .......................................................... 70
  4.2.3 Bayesian Partial Membership Models ....................................... 71
  4.2.4 Exponential Family Factor Analysis ......................................... 72
4.3 Prior to Posterior Analysis .................................................................. 74
  4.3.1 Markov Chain Monte Carlo ..................................................... 74
    Aspects of Implementation ............................................................ 75
4.4 Related Work ............................................................................... 76
4.5 Experimental Results ...................................................................... 77
  4.5.1 Synthetic Binary Data ............................................................. 78
    Noisy Bit Patterns ....................................................................... 78
    Simulated Data from the BPM ....................................................... 80
  4.5.2 Senate Roll Call Data ................................................................ 80
4.6 Discussion .................................................................................... 83
4.7 Conclusion ..................................................................................... 84
References ......................................................................................... 85

In this chapter 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 visualization, and importantly, allow us to make a broad range of insightful probabilistic queries of our data. We emphasize 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.
4.1 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 (Newcomb, 1886; Bishop, 2006), factor analysis (Bartholomew and Knott, 1999), probabilistic principal components analysis (Tipping and Bishop, 1997; Bishop, 2006), mixed membership models (Erosheva et al., 2004), and matrix factorization (Lee and Seung, 1999; Salakhutdinov and Mnih, 2008), 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 modeling: 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; 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 modeling framework to emphasize this wide applicability. In this chapter, we show how mixture models, factor analysis, and partial membership models and their generalization to mixed membership models can be unified under a common modeling framework. Moreover, we show how exponential family likelihoods can be used to provide a very general tool for modeling diverse data types, such as binary, count, or non-negative data, etc. Specifically, we will develop two models: a Bayesian partial membership model (BPM) (Heller et al., 2008) and a Bayesian exponential family factor analysis (EXFA) (Mohamed et al., 2008), and demonstrate the power of these models for accurate prediction and interpretation of data.

As a case study, we will use an 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 U.S. senate and demonstrate the different perspectives of the data that can be obtained, including 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 about 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 4.2.1). 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 4.4). In Section 4.2.3, 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 4.2.4 leads to an exponential family factor analysis. We focus on Markov chain Monte Carlo methods for learning in both models in Section 4.3. 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
factorization (Blei et al., 2003; Erosheva et al., 2004; Mackey et al., 2010) in Section 4.4. We provide some experimental results and explore the roll call data in Section 4.5.

Notation. Throughout this chapter we represent observed data as an $N \times D$ matrix $X = [x_1, \ldots, x_N]^T$, with an individual data point $x_n = [x_{n1}, \ldots, x_{nD}]$. $N$ is the number of data points and $D$ is the number of input features. $\Theta$ is a $K \times D$ matrix of model parameters with rows $\theta_k$. $V$ is a $N \times K$ matrix $V = [v_1, \ldots, v_N]^T$ of latent variables with rows $v_n = [v_{n1}, \ldots, v_{nK}]$, 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.

### 4.2 Membership Models for the Exponential Family

#### 4.2.1 The Exponential Family of Distributions

The choice of likelihood function $p(x|\eta)$ for parameters $\eta$ and observed data $x$ 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 emphasizes the shared properties of many standard distributions, including the binomial, Poisson, gamma, beta, multinomial, and Gaussian distributions (Bickel and Doksum, 2001). 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 modeling framework generalized to all members of the family.

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

$$p(x|\eta) = \exp\{s(x_n)^T \eta + h(x_n) - g(\eta)\}, \quad (4.1)$$

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

Probability distributions that belong to the exponential family also have corresponding conjugate prior distributions $p(\eta)$, for which both $p(\eta)$ and $p(x|\eta)$ have the same functional form. The conjugate prior distribution for the exponential family distribution of Equation (4.1) is:

$$p(\eta) \propto \exp\{\lambda^T \eta - \nu g(\eta) + f(\lambda)\}, \quad (4.2)$$

where $\lambda$ and $\nu$ are hyperparameters of the prior distribution. We use the shorthand $\eta \sim \text{Conj}(\lambda, \nu)$ 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 $\mu$ in $[0,1]$. The exponential family form, using the terms in Equation (4.1), is described using $h(x) = 0$, $\eta = \ln(\frac{\mu}{1-\mu})$ and $g(\eta) = \ln(1 + e^{\nu \eta})$. 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.
4.2.2 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 $x_n$ given parameters $\Theta$ is

$$p(x_n|\Theta) = \sum_{k=1}^{K} \rho_k p_k(x_n|\theta_k),$$  \hspace{1cm} (4.3)

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

$$p(x_n|\Theta) = \sum_{v_n} p(v_n) \prod_{k=1}^{K} (p_k(x_n|\theta_k))^{v_{nk}},$$ \hspace{1cm} (4.4)

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 (4.4) must now be modified, and becomes

$$p(x_n|\Theta) = \int_{v_n} p(v_n) \frac{1}{Z(v_n, \Theta)} \prod_{k=1}^{K} (p_k(x_n|\theta_k))^{v_{nk}} d v_n,$$ \hspace{1cm} (4.5)

where we have integrated over the continuous latent variables rather than summing, and have introduced the normalizing constant $Z$, which is a function of $v_n$ and $\Theta$, to ensure normalization.

By substituting the exponential family distribution (4.1) into Equation (4.5), the likelihood can be expressed as

$$x_n|v_n, \Theta \sim \text{Expon} \left(\sum_k v_{nk} \theta_k\right),$$ \hspace{1cm} (4.6)

which is obtained by combining terms in log-space and requiring the resulting distribution to be normalized. The computation of the normalizing constant $Z$ in Equation (4.5) 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 $\theta_k$ 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:

**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, mixed membership, and topic modeling.

**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) (Bartholomew and Knott, 1999), probabilistic principal components analysis (PCA) (Tipping and Bishop, 1997), and probabilistic matrix factorization (PMF) (Salakhutdinov and Mnih, 2008), 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 nonnegative, allowing for a parts-based explanation of the data (Lee and Seung, 1999).
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 4.1 summarizes this insight and lists some of the models that can arise from this framework.

**TABLE 4.1**
Models which can be derived from the unifying framework for latent variable models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>mixture models</td>
<td></td>
</tr>
<tr>
<td>partial membership models (Heller et al., 2008)</td>
<td>$v_{nk} \in {0, 1}$</td>
</tr>
<tr>
<td>exponential family PCA (Collins et al., 2002; Mohamed et al., 2008)</td>
<td>$v_{nk} \in [0, 1]$</td>
</tr>
<tr>
<td>nonnegative matrix factorization (Lee and Seung, 1999)</td>
<td>$v_{nk} \in \mathbb{R}$</td>
</tr>
<tr>
<td></td>
<td>$v_{nk} \in \mathbb{R}^+$</td>
</tr>
</tbody>
</table>

### 4.2.3 Bayesian Partial Membership Models

We consider a model for partial membership that we refer to as the Bayesian partial membership model (BPM) (Heller et al., 2008). The BPM is a model in which we consider observations have partial membership each of $K$ 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 in between that of being fully Republican or fully Democrat.

Figure 4.1(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. We denote the $K$-dimensional vector of positive hyperparameters by $\alpha$. The generative model is: Draw mixture weights $\rho_k$ from a Dirichlet distribution with

---

**FIGURE 4.1**
Graphical models representing the relationship between latent variables, parameters, and observed data for exponential family latent variable models.
hyperparameters $\alpha$, and a positive scaling factor $a$ from an exponential distribution with hyperparameter $\beta > 0$; then draw a vector of partial memberships $v_n$ from a Dirichlet distribution, representing the extent to which the observation belongs to each of the $K$ clusters.

\[
\rho \sim \text{Dir}(\alpha); \quad a \sim \text{Exp}(\beta), \quad (4.7)
\]

\[
v_n \sim \text{Dir}(a\rho). \quad (4.8)
\]

Each cluster $k$ is characterized by an exponential family distribution with natural parameters $\theta_k$ that are drawn from a conjugate exponential family distribution, with hyperparameters $\lambda$ and $\nu$. Given the latent variables and parameters, each data point is drawn from a data-appropriate exponential family distribution:

\[
\theta_k \sim \text{Conj}(\lambda, \nu), \quad (4.9)
\]

\[
x_n \sim \text{Expon}\left(\sum_k v_{nk}\theta_k\right). \quad (4.10)
\]

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

\[
p(X, \Omega | \Psi) = p(X | V, \Theta) p(V | a, \rho) p(\Theta | \lambda, \nu) p(\rho | \alpha) p(a | \beta)
\]

\[
= \prod_{n=1}^{N} p(x_n | v_n, \Theta) p(v_n | a, \rho) \prod_{k=1}^{K} p(\theta_k | \lambda, \nu) p(\rho | \alpha) p(a | \beta). \quad (4.11)
\]

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

\[
\ln p(X, \Omega | \Psi) = \sum_{n=1}^{N} \left\{ \left( \sum_k v_{nk}\theta_k \right)^{\top} x_n + h(x_n) + g\left( \sum_k v_{nk}\theta_k \right) \right\}
\]

\[
+ \sum_{k=1}^{K} \left[ \lambda^T \theta_k + \nu g(\theta_k) + f(\lambda) \right]
\]

\[
+ N \ln \Gamma \left( \sum_k a\rho_k \right) - N \sum_k \ln \Gamma (a\rho_k) + \sum_n \sum_k (a\rho_k - 1) \ln v_{nk}
\]

\[
+ \ln \Gamma \left( \sum_k \alpha_k \right) - \sum_k \ln \Gamma (\alpha_k) + \sum_k (\alpha_k - 1) \ln \rho_k + \ln b - ba.
\]

We arrive at the BPM model using a continuous latent variable relaxation of the mixture model. As a result, the BPM reduces to mixture modeling when $a \to 0$ with mixing proportions $\rho$, and follows from the limit of Equation (4.8). The BPM bears interesting relationships to several well-known models, including latent Dirichlet allocation (LDA) (Blei et al., 2003), mixed membership models (Erosheva et al., 2004), discrete components analysis (DCA) (Buntine and Jakulin, 2006), and exponential family PCA (Collins et al., 2002; Moustaki and Knott, 2000), which we discuss in Section 4.2.4. Unlike LDA and mixed membership models that capture partial memberships in the form of attribute-specific mixtures, the BPM does not assume a factorization over attributes and provides a general way of combining exponential family distributions with partial membership.

### 4.2.4 Exponential Family Factor Analysis

We now consider a Bayesian model for exponential family factor analysis (EXFA) (Mohamed et al., 2008). We can think of an exponential family factor analysis as a method of decomposing an observed data matrix $X$, which can be of any type supported by the exponential family of distributions,
into two matrices $V$ and $\Theta$; we define the product matrix $P = V\Theta$. Since the likelihood depends only on $V$ and $\Theta$ through their product $P$, this can also be seen as a model for matrix factorization. In traditional factor analysis and probabilistic PCA, the elements of the matrix $P$, which are the means of Gaussian distributions, lie in the same space as that of the data $X$. In the case of EXFA and similar methods for non-Gaussian PCA such as EPCA (Collins et al., 2002; Moustaki and Knott, 2000), 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 4.1(b). Let $m$ and $S$ be hyperparameters representing a $K$-dimensional vector of initial mean values and an initial covariance matrix, respectively. Let $\alpha$ and $\beta$ be the hyperparameters corresponding to the shape and scale parameters of an inverse-gamma distribution. We begin by drawing $\mu$ from a Gaussian distribution and the elements $\sigma_k^2$ of the diagonal matrix $\Sigma$ from an inverse-gamma distribution. For each data point $n$ of the factor score matrix $V$, we draw a $K$-dimensional Gaussian latent variable $v_n$:

$$\mu \sim \mathcal{N}(\mu|m, S); \quad \sigma_k^2 \sim \mathcal{G}(\alpha, \beta)$$

$$v_n \sim \mathcal{N}(v_n|\mu, \Sigma).$$

The data is described by an exponential family distribution with natural parameters given by the product of the latent variables $v_n$ and parameters $\theta_k$. The exponential family distribution modeling the data and the corresponding prior over the model parameters is:

$$\theta_k \sim \text{Conj}(\lambda, \nu)$$

$$x_n|v_n, \Theta \sim \text{Expon}(\sum_k v_{nk}\theta_k).$$

We denote $\Omega = \{V, \Theta, \mu, \Sigma\}$ as the set of unknown parameters with hyperparameters $\Psi = \{m, S, \alpha, \beta, \lambda, \nu\}$. Given this specification, in Equations (4.13)–(4.16), the log joint probability distribution is:

$$p(X, \Omega|\Psi) = p(X|V, \Theta)p(\Theta|\lambda, \nu)p(V|\mu, \Sigma)p(\mu|m, S)p(\Sigma|\alpha, \beta)$$

$$\ln p(X, \Omega|\Psi) = \sum_{n=1}^{N} \left[ \left( \sum_k v_{nk}\theta_k \right)^\top x_n + h(x_n) + g \left( \sum_k v_{nk}\theta_k \right) \right]$$

$$+ \sum_{k=1}^{K} [\lambda^\top \theta_k + \nu g(\theta_k) + f(\lambda)]$$

$$+ \sum_{n=1}^{N} \left[ -\frac{K}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (v_n - \mu)^T \Sigma^{-1} (v_n - \mu) \right]$$

$$- \frac{K}{2} \ln(2\pi) - \frac{1}{2} \ln |S| - \frac{1}{2} (\mu - m)^T S^{-1} (\mu - m)$$

$$+ \sum_{i=1}^{K} \left[ \alpha \ln \beta - \ln \Gamma(\alpha) + (\alpha - 1) \ln \sigma_i^2 - \beta \sigma_i^2 \right],$$

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 cluster, 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.
4.3 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 visualization. To achieve this, we must infer the posterior distribution \( p(\Omega|X, \Psi) \), by which we can visualize 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.

4.3.1 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 (Robert and Casella, 2004; Gilks et al., 1995). For sampling in the models of Sections 4.2.3 and 4.2.4, 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. (1987), 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 \( p(\Omega|\Psi, X) \), with \( \Psi \) being any relevant hyperparameters; we denote \( u \) 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 \( \Omega \) are continuous and differentiable, making it possible to exploit available gradient information.

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

\[
H(\Omega, u) = E(\Omega|\Psi) + K(u), \quad \text{(Hamiltonian Energy)}
\]

\[
E(\Omega|\Psi) = -\ln p(\Omega, X|\Psi), \quad \text{(Potential Energy)}
\]

\[
K(u) = -\frac{1}{2} u^T M u. \quad \text{(Kinetic Energy)}
\]

The Hamiltonian can be seen as the log of an augmented distribution to be sampled from:

\[
p(X, \Omega, u|\Psi) = p(X, \Omega|\Psi)N(u|0, M),
\]

where \( M \) 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 \( \Delta(\Omega) = \frac{\partial E(\Omega)}{\partial \Omega} \). We defer further details of the physical underpinnings describing Hamiltonian dynamics and its appropriateness for MCMC to the work of Neal (2010) and Neal (1993).

We present the full algorithm for HMC in Algorithm 1. Each iteration of HMC has two steps. In the first step, we assume that an initial sample (state) for \( \Omega \) is given and we generate a Gaussian
Evaluate Gradient \( g = \Delta(\theta) \) with initial \( \theta \)\n
\[ g = \text{grad}E(\theta) \]

Evaluate Energy \( E = \mathcal{E}(\theta|\psi) \)\n
\[ E = \text{find}E(\theta) \]

for \( L \) iterations do

Initialize new momentum \( u \) drawn from a Gaussian

\[ K(u) = \frac{1}{2}u^\top u \] and \( H = \mathcal{E}(\theta|\psi) + K(u) \)

\[ \theta^{\text{new}} \leftarrow \theta; \quad g^{\text{new}} \leftarrow g; \]

for \( L \) leapfrog steps do

\[ u \leftarrow u - \frac{\epsilon}{2}g \] \n
[f] Make half-step in \( u \)

\[ \theta^{\text{new}} \leftarrow \theta^{\text{new}} + \epsilon u \]

[f] Make a step in theta

\[ g^{\text{new}} \leftarrow \Delta(\theta^{\text{new}}) \]

[f] gradE(thetaNew)

\[ u \leftarrow u - \frac{\epsilon}{2}g^{\text{new}} \]

[f] make half step in \( u \)

end

\[ E^{\text{new}} = \mathcal{E}(\theta^{\text{new}}|\psi) \]

[f] Enew = findE(thetaNew)

Calculate \( K(u) = \frac{1}{2}u^\top u \)

Hamiltonian \( H^{\text{new}} \leftarrow E^{\text{new}} + K(u) \)

if \( \text{rand}() < \exp(-(H^{\text{new}} - H)) \) then

Accept \( \leftarrow \) True

\[ g \leftarrow g^{\text{new}}; \quad \theta \leftarrow \theta^{\text{new}}; \quad E \leftarrow E^{\text{new}} \]

else

Accept \( \leftarrow \) False

end

end

**Algorithm 1**: Hybrid Monte Carlo (HMC) Sampling (MacKay, 2003).

Variable \( u \) for the momentum (line 4, Algorithm 1). In the second step, we simulate Hamiltonian dynamics, which follow the equations of motion to move the current sample and momentum to a new state. The Hamiltonian dynamics must be discretized for implementation and the most popular discretization is known as the leapfrog method (lines 7–11). The leapfrog approximation is simulated for \( L \) steps using a step-size \( \epsilon \). The samples \( \Omega^* \) and \( u^* \) at the end of the leapfrog steps form the proposed state, which is accepted using the Metropolis criterion (line 15):

\[
\min (1, \exp(-H(\Omega^*, u^*) + H(\Omega, u))).
\]

(4.21)

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

**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 nonnegative or bound between \([0,1]\). We make use of the following transformations:

**BPM with \( a > 0, \sum_k \pi_k = 1 \) and \( \sum_k v_{nk} = 1 \):**
\[ a = \exp(\eta); \quad \pi_k = \frac{\exp(r_k)}{\sum_{k'} \exp(r_{k'})}; \quad v_{nk} = \frac{\exp(\omega_{nk})}{\sum_{k'} \exp(\omega_{nk'})}. \]

**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 \( \epsilon \). In general, the step-size should be chosen to ensure that the sampler’s rejection rate is between 25% and 35%, and to use a large number of leapfrog steps. Here we generally make use of \( L \) between 80 and 100. 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 4.6. Analysis of the optimal acceptance rates for HMC is discussed in Beskos et al. (2010); Neal (2010) provides a great deal of guidance in tuning HMC samplers.

Many datasets 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 \( X = \{ X^{obs}, X^{missing} \} \) and conditioning on the set \( X^{obs} \) 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 1.

### 4.4 Related Work

**Mixed Membership Models and LDA.** In general, mixed membership models (Erosheva et al., 2004) organize the data in two levels using an admixture structure (mixture-of-mixtures model). Latent Dirichlet allocation (LDA) (Blei et al., 2003), as an instance of a mixed membership model, organizes 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 Dirichlet-distributed latent variable \( v \) 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 the most sense when the observations (documents) being modeled 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 is in the chapter by Gruhl and Erosheva (Gruhl and Erosheva, 2013, §2.4) in this volume and complements this discussion.

**Latent Gaussian Models.** EXFA employs a \( K \)-dimensional Gaussian latent variable \( v \) and is thus an example of a latent Gaussian model. This is one of the most established classes of models and includes generalized linear regression models, nonparametric regression using Gaussian processes, state-space and dynamical systems, unsupervised latent variable models such as PCA,
factor analysis (Bartholomew and Knott, 1999), probabilistic matrix factorization (Salakhutdinov and Mnih, 2008), and Gaussian Markov random fields. In generalized linear regression (Bickel and Doksum, 2001), the latent variables $v_n$ are the predictors formed by the product of covariates and regression coefficients; in Gaussian process regression (Rasmussen and Williams, 2006), the latent variables $v$ 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 (Tipping and Bishop, 1997; Bartholomew and Knott, 1999), latent variables $v_n$ are Gaussian with isotropic or diagonal covariances, respectively.

EXFA also follows as a Bayesian interpretation of exponential family PCA (Collins et al., 2002) and generalized latent trait models (Moustaki and Knott, 2000). Instead of fully Bayesian inference, these related models specify an objective function that is optimized 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 realization though, it is easy to see the connection and extension of EXFA to a generalized mixed membership matrix factorization (MMMF) model by instead considering a two-level representation of the data similar to that described by Mackey et al. (2010).

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 nonnegative matrix factorization (NMF) (Lee and Seung, 1999; Buntine and Jakulin, 2006), 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 (Zadeh, 1965; Kosko, 1992). 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 (Bezdek, 1981). Fuzzy $k$-means clustering (Gasch and Eisen, 2002) iteratively minimizes the objective function: $J = \sum_{n} \sum_{k} v_{nk} \gamma f D^2(x_n, c_k)$, 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 $D^2(x_n, c_k)$ is a squared distance between the observation $x_n$ and the cluster centre $c_k$. By varying $\gamma f$, 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 4.5 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 modeling.

### 4.5 Experimental Results

We demonstrate the effectiveness of the models presented in this chapter using synthetic datasets as well as a real-world case study: roll call data from the U.S. 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

4.5.1 Synthetic Binary Data

**Noisy Bit Patterns**

We evaluate the behaviour of EXFA using a synthetic binary dataset. 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 dataset of 300 observations. Noise is then added to the data by flipping bits in the dataset with probability of 0.1 (Tipping, 1999; Mohamed et al., 2008). We use HMC to generate 5000 samples from the EXFA model with \( K = 3 \) factors, and demonstrate the evolution of the sampler in Figure 4.2.

**FIGURE 4.2**

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.

¹See www.shakirm.com/code/EFLVM/.
Since the sampler is initialized 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 4.2. 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 optimization methods (Mohamed et al., 2008).

Plots such as Figure 4.2 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 the number of leapfrog steps that allow us to rapidly reach this region. We fix $L = 80$ and tune $\epsilon$ by monitoring the progression of the sampler.

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 dataset we set 10% of the elements of the matrix as missing, using these elements as a held-out dataset. 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 4.3. 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 complex one, we choose $K = 3$ based 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 4.6.

![Figure 4.3](image-url) Choosing the number of latent factors $K$ by cross-validation. We find that $K = 3$ is an appropriate number of latent factors.
Simulated Data from the BPM

We generated a synthetic binary dataset 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 burn-in. To compare the true partial memberships $V_T$ to the inferred memberships $V_L$, we computed $U_T = V_T V_T^T$ and $U_L = V_L V_L^T$, which is a measure of the degree of shared membership between pairs of observations for the true and inferred partial memberships, respectively (Heller et al., 2008). 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 4.4. The difference between entries of the true and inferred shared memberships $|U_T - U_L|$ 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.

4.5.2 Senate Roll Call Data

Having evaluated the behavior of the BPM and EXFA on synthetic data, we demonstrate their use in exploring membership behavior from the U.S. Senate roll call as a case study. Specifically, we analyze the roll call from the 107th U.S. Congress (2001–2002) (Jakulin, 2002). 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 point (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 values. For the perspective of a political scientist analyzing such data, see the chapter by Gross and Manrique-Vallier (Gross and Manrique-Vallier, 2013).

We analyze the data using the BPM with $K = 2$ clusters, and show results of this analysis in Figure 4.5. 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 4.5 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 authorizing the use of force in Iraq), and Senator Miller, a conservative Democrat who supported George Bush over John Kerry in the 2004 U.S. Presidential election. Lastly, it is interesting to note the location of the outcome data point, which is very much in the
FIGURE 4.5
Analysis of the partial memberships for the 107th U.S. 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.

middle. This makes sense since the 107th Congress was split 50-50 (with Republican Vice President Dick Cheney breaking ties), until Senator Jeffords became an Independent, at which point the Democrats had a one seat majority.

We also analyzed 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 $\gamma_f$, which is typically set by hand. Figure 4.6 shows the change in partial membership for the outcome of the most-allegiant Democrat and Republican senator (using the result of Figure 4.5), 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 $\gamma_f$. 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 two latent factors, can be used to provide a spatial visualization of senators. We show the results of analyzing the roll call data with EXFA in Figure 4.8 using $K = 2$ latent factors, producing 4000 samples from the HMC sampler and using the first half as burn-in. The latent embedding in Figure 4.8 is color-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 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 (Bafumi et al., 2005), which aim to spatially represent political preferences on a left-to-right scale.
As a further comparison of the BPM and EXFA, we also analyze 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 four clusters: one cluster consists solely 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 (Figure 4.7). 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 behavior. 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 modeling flexibility. These two approaches emphasize the tradeoff between modeling efficiency and interpretability that must be considered when analyzing such data.

The BPM gives 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 \( \Theta \Theta^\top \). 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 4.8(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.
4.6 Discussion

Having gained an understanding of exponential family latent variable models and their behavior, we now consider some of the questions that affect our ability to use such models in practice. Questions 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 datasets. 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 modeling assumptions made match our beliefs regarding the process that generated the data, as well as the aim of our modeling effort, whether for visualization, 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 visualization 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 (Erosheva et al., 2007; Gross and Manrique-Vallier, 2013) and mixed membership matrix factorization (Mackey et al., 2010), 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. (2012). Galyardt (2013) 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 (Carlin and Chib, 1995).

**Choosing the Latent Dimensionality.** In Section 4.5 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 (2001) 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 (1999) exploited sparsity by employing automatic relevance determination (ARD), which uses a 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 (2004). More recent approaches have focused on the construction of nonparametric latent factor models using the Indian buffet process or other nonparametric priors to automatically adapt the dimensionality of latent variables (Knowles and Ghahramani, 2010; Bhattacharya and Dunson, 2011).

**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 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 $\epsilon$ (Atchade et al., 2011).

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 (2011). Another way of improving HMC was proposed in Shahbaba et al. (2011), 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. Tuning the HMC parameters can be challenging, especially for the non-expert, and methods now exist for the automatic tuning of HMC’s parameters (Hoffman and Gelman, 2011; Wang et al., 2013). 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 datasets, the availability of scalable inference is an important factor in the practical use of many models. MCMC methods can be shown to scale well to large datasets (Salakhutdinov and Mnih, 2008). 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. (2007), and such an approach can be applied to the BPM. For latent Gaussian models, approximate inference methods such as integrated nested Laplace approximations (INLA) (Rue et al., 2009) 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 (Khan et al., 2010).

## 4.7 Conclusion

In this chapter, we have described a principled Bayesian framework for latent variable modeling that is generalized 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 rep-
resent 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 frame-
work, 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 visualization and predictive tasks and the wide range of
insightful probabilistic queries that can be made using these models.

References

Theory and methods. In Barber, D., Cemgil, A. T., and Chiappa, S. (eds), Bayesian Time Series


291–306.


Bishop, C. M. (1999). Bayesian PCA. In Kearns, M. S., Solla, S. A., and Cohn, D. A. (eds), Advances in
388.

Springer.


