Chapter 2

Graphical models and approximate posterior inference

In this chapter we review latent variable graphical models and exponential families. We discuss variational methods and Gibbs sampling for approximate posterior inference, and derive general forms of these algorithms for a large subclass of models.

2.1 Latent variable graphical models

We use the formalism of *directed graphical models* to describe the independence assumptions of the models developed in the subsequent chapters. A directed graphical model provides a succinct description of the factorization of a joint distribution: *nodes* denote random variables; *edges* denote possible dependence between random variables; and *plates* denote replication of a substructure, with appropriate indexing of the relevant variables.

Graphical models can be used to describe *latent variable models*. Latent variable modeling is a method of developing complicated structured distributions, where the data interact with *latent* or *unobserved* random variables. In the graphical model notation, observed random variables are shaded, and latent random variables are unshaded.

For example, the distribution on the real line in Figure 2.1 (Left) is the *mixture*

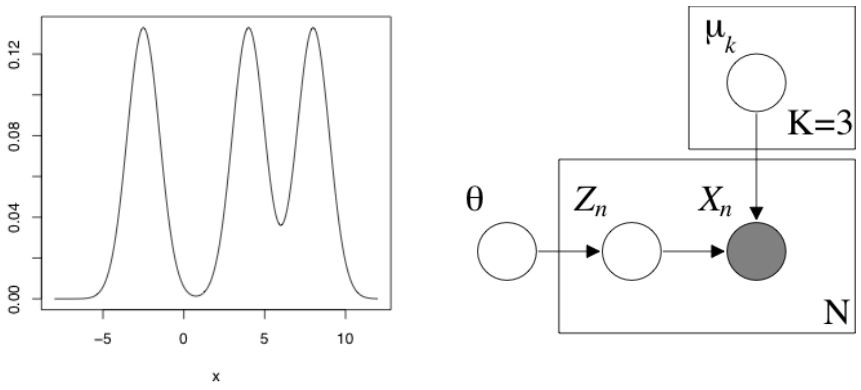

Figure 2.1: (Left) The density of a Gaussian mixture model with three mixture components. (Right) The corresponding graphical model of N data from this density.

distribution formed by combining three unit-variance Gaussian distributions with means $\mu_1 = -2.5$, $\mu_2 = 4$, and $\mu_3 = 8$. A data point is drawn by first choosing a latent variable $Z \in \{1, 2, 3\}$ from a multinomial, and then drawing the data point from $\mathcal{N}(\mu_z, 1)$. This example is illustrated as a graphical model in Figure 2.1 (Right).

The central task of latent variable modeling for data analysis is *posterior inference*, where we determine the distribution of the latent variables conditional on the observations. Loosely, posterior inference can be thought of as a reversal of the generative process which the graphical model illustrates. For example, in the Gaussian mixture with fixed means, we would like to determine the posterior distribution of the indicator Z given a data point x. If $x = 1$, then the posterior $p(Z \mid X = 1, \mu_1, \mu_2, \mu_3)$ is $(0.16, 0.83, 0.01)$.

Traditionally, the structure of the graphical model informs the ease or difficulty of posterior inference. In the models of the subsequent chapters, however, inference is difficult despite a simple graph structure. Thus, we resort to approximate posterior inference, which is the subject of Section 2.2.

Typically, the parameters of the model are not observed (e.g., the means in the Gaussian mixture), and part of the posterior inference problem is to compute their posterior distribution conditional on the data. One option is to adopt the *empirical Bayes* perspective (Morris, 1983; Kass and Steffey, 1989; Maritz and Lwin, 1989),

and find point estimates of the parameters based on maximum likelihood. Such estimates can be found, for example, with the expectation-maximization (EM) algorithm (Dempster et al., 1977), or approximate variant of it (Neal and Hinton, 1999).

Alternatively, we may take a more fully Bayesian approach, placing a prior distribution on the parameters and computing a proper posterior distribution. This is called *hierarchical Bayesian modeling* (Gelman et al., 1995) because it necessitates the specification of a distribution of the parameters, which itself must have parameters called *hyperparameters*.

In a hierarchical Bayesian model, we may still use the empirical Bayes methodology, and find point estimates of the hyperparameters by maximum likelihood. This is often sensible because it affords the advantages of exhibiting uncertainty on the parameters, while avoiding the unpleasant necessity of choosing a fixed hyperparameter or further extending the hierarchy.

2.1.1 Exponential families

All the random variables we will consider are distributed according to *exponential family* distributions. This family of distributions has the form:

$$p(x \,|\, \eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\}, \tag{2.1}$$

where η is the *natural parameter*, $t(x)$ are the *sufficient statistics* for η, and $a(\eta)$ is the *cumulant generating function* or *log partition function*:

$$a(\eta) = \log \int h(x) \exp\{\eta^T t(x)\} dx. \tag{2.2}$$

The derivatives of $a(\eta)$ are the cumulants of $t(x)$. In particular, the first two derivatives are:

$$a'(\eta) = \mathrm{E}_\eta \left[t(X) \right] \tag{2.3}$$

$$a''(\eta) = \mathrm{Var}_\eta \left[t(X) \right]. \tag{2.4}$$

The functions $a(\eta)$ and $h(x)$ are determined by the form and dimension of $t(x)$. For example, if x is real valued and $t(x) = (x, x^2)$, then the corresponding exponential

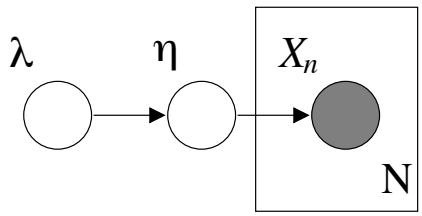

Figure 2.2: Graphical model representation of iid data $X_{1:N}$ from $p(x \mid \eta)$, where η is itself distributed by $p(\eta \mid \lambda)$ for a fixed hyperparameter λ. Computation in this model is facilitated when X_n is in the exponential family, and η is distributed by the conjugate prior.

family is Gaussian. If $t(x)$ is a multidimensional vector with all zeros and a single one, then the corresponding exponential family distribution is multinomial. An exponential family for positive reals is the Gamma distribution, and an exponential family for positive integers is the Poisson distribution.

See Brown (1986) for a thorough analysis of the properties of exponential family distributions.

2.1.2 Conjugate exponential families

In a hierarchical Bayesian model, we must specify a prior distribution of the parameters. In this section, we describe a family of priors which facilitate computations in such a model.

Let X be a random variable distributed according to an exponential family with natural parameter η and log normalizer $a(\eta)$. A *conjugate prior* of η, with natural parameter λ, has the form:

$$p(\eta \mid \lambda) = h(\eta) \exp\{\lambda_1^T \eta + \lambda_2(-a(\eta)) - a(\lambda)\}.$$

The parameter λ has dimension $\dim(\eta) + 1$ and the sufficient statistic is $t(\eta) = (\eta, -a(\eta))$. We decompose $lambda = (\lambda_1, \lambda_2)$ such that λ_1 contains the first $\dim(\eta)$ components and λ_2 is a scalar. (Note that we overload $a(\cdot)$ to be the log normalizer for the parameter in the argument.)

The conjugate distribution is a convenient choice of prior, because the corresponding posterior will have the same form. Consider the simple model illustrated in Figure 2.2 where $X_{1:N}$ are independent and identically distributed (iid) variables from the exponential family distribution $p(x_n \,|\, \eta)$, and $p(\eta \,|\, \lambda)$ is the conjugate prior. The posterior of η is:

$$\begin{aligned} p(\eta \,|\, \lambda, x_{1:N}) &\propto p(\eta \,|\, \lambda) p(x_{1:N} \,|\, \eta) \\ &\propto h(\eta) \exp\{\lambda_1^T \eta + \lambda_2(-a(\eta))\} \prod_{n=1}^{N} \exp\{\eta^T t(x_n) - a(\eta)\} \\ &= h(\eta) \exp\{(\lambda_1 + \textstyle\sum_{n=1}^{N} t(x_n))^T \eta + (\lambda_2 + N)(-a(\eta))\}, \end{aligned} \tag{2.5}$$

which is the same type of distribution as $p(\eta \,|\, \lambda)$, with posterior parameters $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2)$:

$$\begin{aligned} \hat{\lambda}_1 &= \lambda_1 + \textstyle\sum_{n=1}^{N} t(x_n) \\ \hat{\lambda}_2 &= \lambda_2 + N. \end{aligned} \tag{2.6}$$

The posterior, conditional on any amount of data, can be fully specified by the prior parameters, the sum of the sufficient statistics, and the number of data points.

A second convenience of the conjugate prior is for computing the marginal distribution $p(x \,|\, \lambda) = \int p(x \,|\, \eta) p(\eta \,|\, \lambda) d\eta$. If $p(\eta \,|\, \lambda)$ is conjugate, then:

$$\begin{aligned} p(x \,|\, \lambda) &= h(x) \int \exp\{\eta^T t(x) - a(\eta)\} h(\eta) \exp\{\lambda_1^T \eta + \lambda_2(-a(\eta)) - a(\lambda)\} d\eta \\ &= h(x) \int h(\eta) \exp\{(\lambda_1 + t(x))^T \eta + (\lambda_2 + 1)(-a(\eta))\} d\eta \exp\{-a(\lambda)\} \\ &= h(x) \exp\{a((\lambda_1 + t(x), \lambda_2 + 1)) - a(\lambda)\}. \end{aligned} \tag{2.7}$$

Thus, if the log normalizer is easy to compute then the marginal distribution will also be easy to compute.

Finally, the conjugate prior facilitates computing the *predictive distribution* $p(x \,|\, x_{1:N}, \lambda)$, which is simply a marginal under the posterior parameters in Eq. (2.6).

Example: Gaussian with Gaussian prior on the mean

Suppose the data are real vectors distributed according to a Gaussian distribution with fixed inverse covariance Λ. The exponential family form of the data density is:

$$p(x \,|\, \eta) = \exp\left\{-\frac{1}{2}(d\log 2\pi - \log|\Lambda| + \eta^T\Lambda^{-1}\eta) + \eta^T x - \frac{x^T\Lambda x}{2}\right\},$$

where:

$$\begin{aligned} h(x) &= \exp\left\{-\frac{1}{2}(d\log 2\pi - \log\Lambda)\right\} \\ a(\eta) &= -\eta^T\Lambda^{-1}\eta. \end{aligned}$$

The conjugate prior is thus of the form:

$$p(\eta \,|\, \lambda) \propto \exp\left\{\lambda_1^T\eta - \lambda_2\left(\frac{\eta^T\Lambda^{-1}\eta}{2}\right)\right\},$$

which is a Gaussian with natural parameters λ_1 and $\lambda_2\Lambda^{-1}$. Note that its covariance is the scaled inverse covariance of the data $\frac{\Lambda}{\lambda_2}$. The log normalization is:

$$a(\lambda) = -\frac{1}{2}\log|\lambda_2\Lambda^{-1}| + \frac{\lambda_1^T\Lambda\lambda_1}{\lambda_2},$$

from which we can compute the expected sufficient statistics of η:

$$\begin{aligned} \mathrm{E}[\eta] &= \frac{(\Lambda + \Lambda^T)\lambda_1}{\lambda_2} \\ \mathrm{E}[-a(\eta)] &= \frac{d}{\lambda_2} - \frac{\lambda_1^T\Lambda\lambda_1}{\lambda_2^2}. \end{aligned}$$

2.1.3 Exponential family conditionals

Conditional on all the other variables in a directed graphical model, the distribution of a particular variable depends only on its *Markov blanket*, which is the set containing its parents, children, and other parents of its children. To facilitate approximate posterior inference, we consider models for which the conditional distribution of every node given its Markov blanket is in an exponential family.

One possible substructure which meets this requirement is the conjugate-exponential family model of Figure 2.2. Conditional on η, the distribution of X_n is in an exponential family. Moreover, as we have shown above, the conditional distribution of $\eta \,|\, \{\lambda, x_{1:N}\}$ is also in an exponential family.

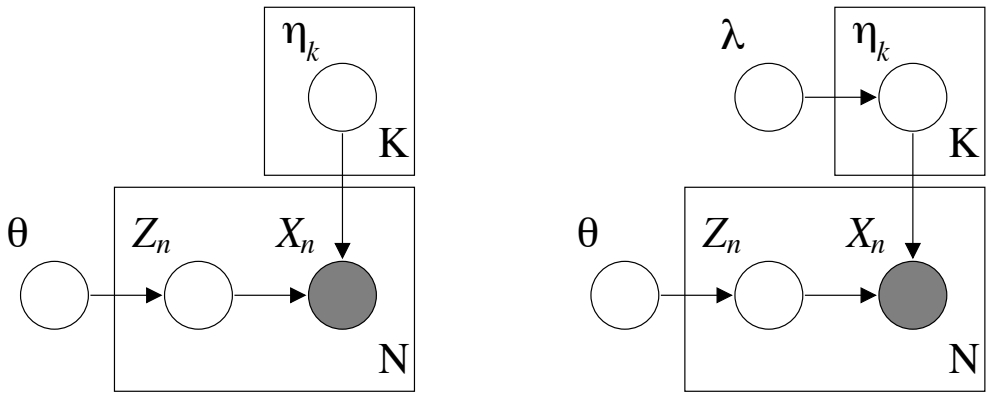

Figure 2.3: (Left) Graphical model representation of a K-mixture model. (Right) A Bayesian K-mixture model.

A second possibility is for the distribution of a variable to be a *mixture* of exponential family distributions. This important substructure is illustrated in Figure 2.3 (Left), where $\eta_{1:K}$ are exponential family parameters and θ is a K-dimensional multinomial parameter. The variables $X_{1:N}$ can be thought of as drawn from a two-stage generative process: first, choose Z_n from Mult(θ); then, choose X_n from the distribution indexed by that value $p(x_n \mid \eta_{z_n})$.

Note that we represent multinomial variables using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, using superscripts to denote components, the kth item is represented by a K-vector z such that $z^k = 1$ and $z^\ell = 0$ for $\ell \neq k$.

We confirm the conditional distributions of nodes X_n and Z_n, given their respective Markov blankets, are in the exponential family. First, by definition, the conditional distribution $p(x_n \mid z_n)$ is a member of the η-indexed exponential family. Second, the conditional distribution $p(z_n \mid x_n)$ is a multinomial:

$$p(z_n \mid x_n, \theta, \eta_{1:K}) \propto p(z_n \mid \theta) p(x_n \mid z_n, \eta_{1:K}),$$

which is also in the exponential family.

In the hierarchical mixture model of Figure 2.3 (Right), we can place the conjugate prior on $\eta_{1:K}$. The distribution of $\eta_k \mid \{z_{1:N}, x_{1:N}\}$ remains in the exponential family as a consequence of the analysis in Eq. (2.5). In particular, we condition only on

those x_n for which $z_n^k = 1$.

By combining the substructures described above, we can build complicated families of distributions which satisfy the requirement that each node, conditional on its Markov blanket, is distributed according to an exponential family distribution. This collection of families contains many probabilistic models, including Markov random fields, Kalman filters, hidden Markov models, mixture models, and hierarchical Bayesian models with conjugate and mixture of conjugate priors.

2.2 Approximate posterior inference

The central computational challenge in latent variable modeling is to compute the posterior distribution of the latent variables conditional on some observations. Except in rudimentary models, such as Figures 2.2 and 2.3, exact posterior inference is intractable and practical data analysis relies on good approximate alternatives. In this section, we describe two general techniques for the class of graphical models which satisfy the conditional exponential family restriction described above.

In the following, we consider a latent variable probabilistic model with parameters η, observed variables $\mathbf{x} = x_{1:N}$ and latent variables $\mathbf{Z} = Z_{1:M}$. The posterior distribution of the latent variables is:

$$p(z_{1:M} \,|\, x_{1:N}, \eta) = \frac{p(x_{1:N}, z_{1:M} \,|\, \eta)}{\int p(x_{1:N}, z_{1:M} \,|\, \eta) dz_{1:M}}.$$

Under the assumptions, the numerator is in the exponential family and should be easy to compute. The denominator, however, is often intractable due to the nature of $z_{1:M}$. For example, if the latent variables are realizations of one of K values, then this integral is a sum over K^M possibilities. (E.g., this is true for the hierarchical mixture model of Figure 2.3 Right.)

2.2.1 Gibbs sampling

Markov chain Monte Carlo (MCMC) sampling is the most widely used method of approximate inference. The idea behind MCMC is to approximate a distribution by

forming an empirical estimate from samples. We construct a Markov chain with the appropriate stationary distribution, and collect the samples from a chain which has converged or "burned in".

The simplest MCMC algorithm is the *Gibbs sampler*, in which the Markov chain is defined by iteratively sampling each variable conditional on the most recently sampled values of the other variables. This is a form of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), and thus yields a chain with the desired stationary distribution (Geman and Geman, 1984; Gelfand and Smith, 1990; Neal, 1993).

In approximate posterior inference, the distribution of interest is the posterior $p(\mathbf{z} \mid \mathbf{x}, \eta)$. Thus, an iteration of the Gibbs sampler draws each latent variable z_i from $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}, \eta)$. After the resulting chain has converged, we collect B samples $\{\mathbf{z}_1, \ldots, \mathbf{z}_B\}$ and approximate the posterior with an empirical distribution:

$$p(\mathbf{z} \mid \mathbf{x}, \eta) = \frac{1}{B} \sum_{b=1}^{B} \delta_{\mathbf{z}_b}(\mathbf{z}).$$

This use of Gibbs sampling has revolutionized hierarchical Bayesian modeling (Gelfand and Smith, 1990).

In the models described in Section 2.1.3, the every variable, conditional on its Markov blanket, is distributed according to an exponential family distribution. Gibbs sampling in this setting is thus straightforward, provided we can easily compute the conditional exponential family parameter for each variable.[1]

2.2.2 Mean-field variational methods

Variational inference provides an alternative, deterministic methodology for approximating likelihoods and posteriors in an intractable probabilistic model (Wainwright and Jordan, 2003). We first review the basic idea in the context of the exponential family of distributions, and then turn to its application to approximating a posterior.

[1] In fact, Gibbs sampling is so straightforward in this case, that one can automatically generate Gibbs sampling code from a graph structure and parameterization using the popular BUGS package (Gilks et al., 1996).

Consider the η-indexed exponential family distribution in Eq. (**??**) and recall the cumulant generating function $a(\eta)$:

$$a(\eta) = \log \int \exp\{\eta^T t(z)\} h(z) dz.$$

As discussed by Wainwright and Jordan (2003), this quantity can be expressed variationally as:

$$a(\eta) = \sup_{\mu \in \mathcal{M}} \{\eta^T \mu - a^*(\mu)\}, \tag{2.8}$$

where $a^*(\mu)$ is the Fenchel-Legendre conjugate of $a(\eta)$ (Rockafellar, 1970), and $\mathcal{M}$ is the set of *realizable expected sufficient statistics*: $\mathcal{M} = \{\mu : \mu = \int t(z)p(z)h(z)dz, \text{for some } p\}$. There is a one-to-one mapping between parameters η and the interior of $\mathcal{M}$ (Brown, 1986). Accordingly, the interior of $\mathcal{M}$ is often referred to as the set of *mean parameters*.

Let $\eta(\mu)$ be a natural parameter corresponding to the mean parameter $\mu \in \mathcal{M}$; thus $\mathrm{E}_\eta [t(Z)] = \mu$. Let $q(z \,|\, \eta(\mu))$ denote the corresponding density. Given $\mu \in \mathcal{M}$, a short calculation shows that $a^*(\mu)$ is the negative entropy of q:

$$a^*(\mu) = \mathrm{E}_{\eta(\mu)} \left[\log q(Z \,|\, \eta(\mu))\right]. \tag{2.9}$$

Given its definition as a Fenchel conjugate, the negative entropy is convex.

In many models of interest, $a(\eta)$ is not feasible to compute because of the complexity of $\mathcal{M}$ or the lack of any explicit form for $a^*(\mu)$. However, we can bound $a(\eta)$ using Eq. (2.8):

$$a(\eta) \geq \mu^T \eta - a^*(\mu), \tag{2.10}$$

for any mean parameter $\mu \in \mathcal{M}$. Moreover, the tightness of the bound is measured by a Kullback-Leibler divergence expressed in terms of a mixed parameterization:

$$\begin{aligned} D(q(z \,|\, \eta(\mu)) \;||\; p(z \,|\, \eta)) &= \mathrm{E}_{\eta(\mu)} \left[\log q(z \,|\, \eta(\mu)) - \log p(z \,|\, \eta)\right] \\ &= \eta(\mu)^T \mu - a(\eta(\mu)) - \eta^T \mu + a(\eta) \\ &= a(\eta) - \eta^T \mu + a^*(\eta(\mu)). \end{aligned} \tag{2.11}$$

Mean-field variational methods are a special class of variational methods that are based on maximizing the bound in Eq. (2.10) with respect to a subset $\mathcal{M}_{\text{tract}}$

of the space $\mathcal{M}$ of realizable mean parameters. In particular, $\mathcal{M}_{\text{tract}}$ is chosen so that $a^*(\eta(\mu))$ can be evaluated tractably and so that the maximization over $\mathcal{M}_{\text{tract}}$ can be performed tractably. Equivalently, given the result in Eq. (2.11), mean-field variational methods minimize the KL divergence $D(q(z \,|\, \eta(\mu)) \;||\; p(z \,|\, \eta))$ with respect to its first argument.

If the distribution of interest is a posterior, then $a(\eta)$ is the log likelihood. Consider in particular a latent variable probabilistic model with hyperparameters η, observed variables $\mathbf{x} = \{x_1, \ldots, x_N\}$, and latent variables $\mathbf{z} = \{z_1, \ldots, z_M\}$. The posterior can be written as:

$$p(\mathbf{z} \,|\, \mathbf{x}, \eta) = \exp\{\log p(\mathbf{z}, \mathbf{x} \,|\, \eta) - \log p(\mathbf{x} \,|\, \eta)\}, \tag{2.12}$$

and the bound in Eq. (2.10) applies directly. We have:

$$\log p(\mathbf{x} \,|\, \eta) \geq \mathrm{E}_q \left[\log p(\mathbf{x}, \mathbf{Z} \,|\, \eta)\right] - \mathrm{E}_q \left[\log q(\mathbf{Z})\right]. \tag{2.13}$$

This equation holds for any q via Jensen's inequality, but, as our analysis has shown, it is useful specifically for q of the form $q(z \,|\, \eta(\mu))$ for $\mu \in \mathcal{M}_{\text{tract}}$.

A straightforward way to construct tractable subfamilies of exponential family distributions is to consider factorized families, in which each factor is an exponential family distribution depending on a so-called *variational parameter*. In particular, let us consider distributions of the form $q(\mathbf{z} \,|\, \boldsymbol{\nu}) = \prod_{i=1}^{M} q(z_i \,|\, \nu_i)$, where $\boldsymbol{\nu} = \{\nu_1, \nu_2, ..., \nu_M\}$ are variational parameters. Using this class of distributions, we simplify the likelihood bound using the chain rule:

$$\log p(\mathbf{x} \,|\, \eta) \geq \log p(\mathbf{x} \,|\, \eta) + \sum_{m=1}^{M} \mathrm{E}_q \left[\log p(Z_m \,|\, \mathbf{x}, Z_1, \ldots, Z_{m-1}, \eta)\right] - \sum_{m=1}^{M} \mathrm{E}_q \left[\log q(Z_m \,|\, \nu_m)\right]. \tag{2.14}$$

To obtain the best approximation available within the factorized subfamily, we now wish to optimize this expression with respect to ν_i.

To optimize with respect to ν_i, reorder $\mathbf{z}$ such that z_i is last in the list. The portion of Eq. (2.14) depending on ν_i is:

$$\ell_i = \mathrm{E}_q \left[\log p(z_i \,|\, \mathbf{z}_{-i}, \mathbf{x}, \eta)\right] - \mathrm{E}_q \left[\log q(z_i \,|\, \nu_i)\right]. \tag{2.15}$$

Given our assumption that the variational distribution $q(z_i \mid \nu_i)$ is in the exponential family, we have:

$$q(z_i \mid \nu_i) = h(z_i) \exp\{\nu_i^T z_i - a(\nu_i)\},$$

and Eq. (2.15) simplifies as follows:

$$\begin{aligned} \ell_i &= \mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta) - \log h(Z_i) - \nu_i^T Z_i + a(\nu_i)\right] \\ &= \mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta)\right] - \mathrm{E}_q \left[\log h(Z_i)\right] - \nu_i^T a'(\nu_i) + a(\nu_i), \end{aligned}$$

because $\mathrm{E}_q [Z_i] = a'(\nu_i)$. The derivative with respect to ν_i is:

$$\frac{\partial}{\partial \nu_i} \ell_i = \frac{\partial}{\partial \nu_i} \left(\mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta)\right] - \mathrm{E}_q \left[\log h(Z_i)\right]\right) - \nu_i^T a''(\nu_i). \tag{2.16}$$

Thus the optimal ν_i satisfies:

$$\nu_i = \left[a''(\nu_i)\right]^{-1} \left(\frac{\partial}{\partial \nu_i} \mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta)\right] - \frac{\partial}{\partial \nu_i} \mathrm{E}_q \left[\log h(Z_i)\right]\right). \tag{2.17}$$

The result in Eq. (2.17) is general. Under the assumptions of Section 2.1.3, a further simplification is achieved. In particular, when the conditional $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}, \eta)$ is an exponential family distribution:

$$p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}, \eta) = h(z_i) \exp\{g_i(\mathbf{z}_{-i}, \mathbf{x}, \eta)^T z_i - a(g_i(\mathbf{z}_{-i}, \mathbf{x}, \eta))\},$$

where $g_i(\mathbf{z}_{-i}, \mathbf{x}, \eta)$ denotes the natural parameter for z_i when conditioning on the remaining latent variables and the observations. This yields simplified expressions for the expected log probability of Z_i and its first derivative:

$$\begin{aligned} \mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta)\right] &= \mathrm{E} \left[\log h(Z_i)\right] + \mathrm{E}_q \left[g_i(\mathbf{Z}_{-i}, \mathbf{x}, \eta)\right]^T a'(\nu_i) - \mathrm{E}_q \left[a(g_i(\mathbf{Z}_{-i}, \mathbf{x}, \eta))\right] \\ \frac{\partial}{\partial \nu_i} \mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta)\right] &= \frac{\partial}{\partial \nu_i} \mathrm{E}_q \left[\log h(Z_i)\right] + \mathrm{E}_q \left[g_i(\mathbf{Z}_{-i}, \mathbf{x}, \eta)\right]^T a''(\nu_i). \end{aligned}$$

Using the first derivative in Eq. (2.17), the maximum is attained at:

$$\nu_i = \mathrm{E}_q \left[g_i(\mathbf{Z}_{-i}, \mathbf{x}, \eta)\right]. \tag{2.18}$$

We define a coordinate ascent algorithm based on Eq. (2.18) by iteratively updating ν_i for $i \in \{1, \ldots, N\}$. Such an algorithm finds a local maximum of Eq. (2.13) by

Proposition 2.7.1 of Bertsekas (1999), under the condition that the right-hand side of Eq. (2.15) is strictly convex. Further perspectives on algorithms of this kind can be found in Xing et al. (2003) and Beal (2003).

Notice the interesting relationship of this algorithm to the Gibbs sampler. In Gibbs sampling, we iteratively draw the latent variables z_i from the distribution $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}, \eta)$. In mean-field variational inference, we iteratively update the variational parameters of z_i to be equal to the expected value of the parameter g_i of the conditional distribution $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}, \eta)$, where the expectation is taken under the variational distribution.[2]

2.3 Discussion

In this chapter, we described the directed graphical model formalism, and used it to represent latent variable models for data analysis. For the class of models with conditional exponential family distributions—for which conjugate priors and mixture distributions are sufficient—we derived Gibbs sampling and mean-field variational methods of approximate posterior inference.

Choosing an approximate inference technique is an important part of the data analysis process. In this thesis, we typically prefer mean-field variational methods to Gibbs sampling. Gibbs sampling does have some advantages over variational inference. It gives samples from the exact posterior, while estimates based on variational methods incur an unknown bias. However, obtaining correct samples crucially depends on the Markov chain's convergence to its stationary distribution. This can be a slow process, and assessing whether the chain has converged is difficult. Theoretical bounds on the mixing time are of little practical use, and there is no consensus on how to choose one of the several empirical methods developed for this purpose (**?**).

On the other hand, variational methods are deterministic and have a clear convergence criterion given by the bound in Eq. (2.13). Furthermore, they are typically

[2]This relationship has inspired the software package VIBES (Bishop et al., 2003), which is a variational version of the BUGS package (Gilks et al., 1996).

faster than Gibbs sampling, as we will demonstrate empirically in Section 5.5. This is particularly important in view of the goal of efficient data analysis of large collections of text and images.