Bayesian parameter estimation via variational methods

Tommi S. Jaakkola
Massachusetts Institute of Technology
Cambridge, MA

Michael I. Jordan
University of California
Berkeley, CA

November 4, 1999

Abstract

We consider a logistic regression model with a Gaussian prior distribution over the parameters. We show that an accurate variational transformation can be used to obtain a closed form approximation to the posterior distribution of the parameters thereby yielding an approximate posterior predictive model. This approach is readily extended to binary graphical model with complete observations. For graphical models with incomplete observations we utilize an additional variational transformation and again obtain a closed form approximation to the posterior. Finally, we show that the dual of the regression problem gives a latent variable density model, the variational formulation of which leads to exactly solvable EM updates.

1 Introduction

Bayesian methods have a number of virtues, particularly their uniform treatment of uncertainty at all levels of the modeling process. The formalism also allows ready incorporation of prior knowledge and the seamless combination of such knowledge with observed data (Bernardo & Smith 1994, Gelman 1995, Heckerman et al. 1995). The elegant semantics, however, often comes at a sizable computational cost—posterior distributions resulting from the incorporation of observed data must be represented and updated, and this generally involves high-dimensional integration. The computational cost involved in carrying out these operations can call into question the viability of Bayesian methods even in relatively simple settings, such as generalized linear models (McCullagh & Nelder 1983). We concern ourselves in this paper with a particular generalized linear model—logistic regression—and we focus on Bayesian calculations that are computationally tractable. In particular we describe a flexible deterministic approximation procedure that allows the posterior distribution in logistic regression to be represented and updated efficiently. We also show how our methods permit a Bayesian treatment of a more complex model—a directed graphical model (a "belief network") in which each node is a logistic regression model.

The deterministic approximation methods that we develop in this paper are known generically as *variational methods*. Variational techniques have been used extensively in the physics literature (see, e.g., Parisi 1988, Sakurai 1985) and have also found applications in statistics (Rustagi 1976). Roughly speaking, the objective of these methods is to transform the problem of interest into an optimization problem via the introduction of extra degrees of freedom known as *variational parameters*. For fixed values of the variational parameters the transformed problem often has a closed form solution, providing an approximate solution to the original problem. The variational parameters are adjusted via an optimization algorithm to yield an improving sequence of approximations. For an introduction to variational methods in the context of graphical models see Jordan et al. (1999).

Let us briefly sketch the variational method that we develop in this paper. We study a logistic regression model with a Gaussian prior on the parameter vector. Our variational transformation replaces the logistic function with an adjustable lower bound that has a Gaussian form; that is, an exponential of a quadratic function of the parameters. The product of the prior and the variationally transformed likelihood thus yields a Gaussian expression for the posterior (conjugacy), which we optimize variationally. This procedure is iterated for each successive data point.

Our methods can be compared to the Laplace approximation for logistic regression (cf. Spiegelhalter & Lauritzen 1990), a closely related method which also utilizes a Gaussian approximation to the posterior. To anticipate the discussion in following sections, we will see that the variational approach has an advantage over the Laplace approximation; in particular, the use of variational parameters gives the variational approach greater flexibility. We will show that this flexibility translates into improved accuracy of the approximation.

Variational methods can also be contrasted with sampling techniques, which have become the method of choice in Bayesian statistics (Thomas et al. 1992, Neal 1993, Gilks et al. 1996). Sampling techniques enjoy wide applicability and can be powerful in evaluating multi-dimensional integrals and representing posterior distributions. They do not, however, yield closed form solutions nor do they guarantee monotonically improving approximations. It is precisely these features that characterize variational methods.

The paper is organized as follows. First we describe in some detail a variational approximation method for Bayesian logistic regression. This is followed by an evaluation of the accuracy of the method and a comparison to Laplace approximation. We then extend the framework to belief networks, considering both complete data and incomplete data. Finally, we consider the dual of the regression problem and show that our techniques lead to exactly solvable EM updates.

2 Bayesian logistic regression

We begin with a logistic regression model given by:

$$P(S=1|X,\theta) = g\left(\theta^T X\right), \tag{1}$$

where $g(x) = (1+e^{-x})^{-1}$ is the logistic function, S the binary response variable, and $X = \{X_1, \ldots, X_r\}$ the set of explanatory variables. We represent the uncertainty in the parameter values θ via a prior distribution $P(\theta)$ which we assume to be a Gaussian with possibly full covariance structure. Our predictive distribution is therefore:

$$P(S|X) = \int P(S|X,\theta)P(\theta)d\theta. \tag{2}$$

In order to utilize this distribution we need to be able to compute the posterior parameter distribution $P(\theta|D^1, \ldots, D^T)$, where we assume that each $D^t = \{S^t, X_1^t, \ldots, X_r^t\}$ is a complete observation. This calculation is intractable for large n or T, thus we consider a variational approximation.

Our approach involves finding a variational transformation of the logistic function and using this transformed function as an approximate likelihood.

In particular we wish to consider transformations that combine readily with a Gaussian prior in the sense that the Gaussian prior becomes the conjugate prior to the transformed likelihood. We begin by introducing the type of variational transformations we will use for this purpose.

2.1 A brief introduction to variational methods

Consider any continuously differentiable *convex* function $f(z)$. Figure 1 provides an example of a convex function that we will make use of later on. Convexity of this function guarantees by definition that any tangent line always remains below the function itself. We may thus interpret the collection of all the tangent lines as a parameterized family of lower bounds for this convex function (cf. convex duality, Rockafellar 1976). The tangents in this family are naturally parameterized by their locations. From the point of view of approximating the convex non-linear function f, it seems natural to use one of the simpler tangent lines as a lower bound. To formulate this a little more precisely, let $L(z; z_0)$ be the tangent line at $z = z_0$,

$$L(z; z_0) = f(z_0) + \frac{\partial}{\partial z} f(z)_{|z=z_0} (z - z_0), \tag{3}$$

then it follows that $f(z) \geq L(z; z_0)$ for all z, z_0 and $f(z_0) = L(z_0; z_0)$. In the terminology of variational methods, $L(z; z_0)$ is a variational lower bound of $f(z)$ where the parameter z_0 is known as the *variational parameter*. Since the lower bound $L(z; z_0)$ is considerably simpler (linear in this case) than the non-linear function $f(z)$, it may be attractive to substitute the lower bound for f. Note that we are free to adjust the variational parameter z_0, the location of the tangent, so as to make $L(z; z_0)$ as accurate an approximation of $f(z)$ as possible around the point of interest, i.e., when $z \approx z_0$. The quality of this approximation degrades as z receeds from z_0; the rate at which this happens depends on the curvature of $f(z)$. Whenever the function f has relatively low curvature as is the case in Figure 1, the adjustable linear approximation seems quite attractive.

2.2 Variational methods in Bayesian logistic regression

Here we illustrate how variational methods, of the type described above, can be used to transform the logistic likelihood function into a form that

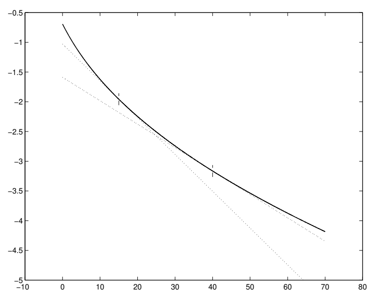


Figure 1: A convex function f and its two tangent lines. The locations of the tangents are indicated with short vertical line segments.

readily combines with the Gaussian prior (conjugacy). More precisely, the transformed logistic function should depend on the parameters θ at most quadratically in the exponent. We begin by symmetrizing the log logistic function:

$$\log g(x) = -\log(1 + e^{-x}) = x/2 - \log(e^{x/2} + e^{-x/2}), \tag{4}$$

and noting that $f(x) = -\log(e^{x/2} + e^{-x/2})$, is a convex function in the variable x^2. (This is readily verified by taking second derivatives; the behavior of $f(x)$ as a function of x^2 is shown in Figure 1). As discussed above, a tangent surface to a convex function is a global lower bound for the function and thus we can bound $f(x)$ globally with a first order Taylor expansion in the variable x^2:

$$\begin{aligned} f(x) &\geq f(\xi) + \frac{\partial f(\xi)}{\partial(\xi^2)}(x^2 - \xi^2) &(5)\\ &= -\xi/2 + \log g(\xi) + \frac{1}{4\xi}\tanh(\xi/2)(x^2 - \xi^2). &(6)\end{aligned}$$

Note that this lower bound is exact whenever $\xi^2 = x^2$. Combining this result with Eq. (4) and exponentiating yields the desired variational transformation of the logistic function:

$$P(S|X,\theta) = g(H_S) \geq g(\xi)\ \exp\left\{(H_S - \xi)/2 - \lambda(\xi)(H_S^2 - \xi^2)\right\}, \tag{7}$$

where $H_S = (2S-1)\theta^T X$ and $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$. We also introduce the following notation:

$$\underline{P}(S|X,\theta,\xi) \equiv g(\xi)\ \exp\left\{(H_S - \xi)/2 - \lambda(\xi)(H_S^2 - \xi^2)\right\}; \tag{8}$$

that is, $\underline{P}(S|X,\theta,\xi)$ denotes the variational lower bound on the logistic function $g(H_S)$. As a lower bound it is no longer normalized. We refer to eq. (8) as a *ξ-transformation* of the conditional probability.

For each fixed value of H_S we can in fact recover the exact value of the logistic function via a particular choice of the variational parameter. Indeed, maximizing the lower bound with respect to ξ yields $\xi = H_S$; substituting this value back into the lower bound recovers the original conditional probability. For all other values of ξ we obtain a lower bound.

The true posterior $P(\theta|D)$ can be computed by normalizing $P(S|X,\theta)P(\theta)$. Given that this calculation is not feasible in general, we instead form the bound:

$$P(S|X,\theta)P(\theta) \geq \underline{P}(S|X,\theta,\xi)P(\theta) \tag{9}$$

and normalize the variational approximation $\underline{P}(S|X,\theta,\xi)P(\theta)$. Given that $P(\theta)$ is Gaussian and given our choice of a Gaussian variational form for $\underline{P}(S|X,\theta,\xi)$, the normalized variational distribution is a Gaussian. Note that although $\underline{P}(S|X,\theta,\xi)$ is a lower bound on the true conditional probability, our variational posterior approximation is a proper density and thus no longer a bound. This approximate Bayesian update amounts to updating the prior mean μ and the prior covariance matrix Σ into the posterior mean and the posterior covariance matrix. Omitting the algebra we find that the updates take the following form:

$$\Sigma_{pos}^{-1} = \Sigma^{-1} + 2\lambda(\xi)\, XX^T \tag{10}$$

$$\mu_{pos} = \Sigma_{pos}\left[\Sigma^{-1}\mu + (S - 1/2)X\right] \tag{11}$$

for a single observation (S, X), where $X = [X_1 \ldots X_r]^T$. Successive observations can be incorporated into the posterior by applying these updates recursively.

Our work is not finished, however, because the posterior covariance matrix depends on the variational parameter ξ through $\lambda(\xi)$ and we have yet to specify ξ. We choose ξ via an optimization procedure; in particular, we find a value of ξ that yields a tight lower bound in eq. (9). The fact that the variational expression in eq. (9) is a lower bound is important—it allows us to use the EM algorithm to perform the optimization. We derive such an EM algorithm in Appendix A; the result is the following (closed form) update equation for ξ:

$$\xi^2 = E\left\{(\theta^T X)^2\right\} = X^T \Sigma_{post}\, X + (X^T \mu_{post})^2, \tag{12}$$

where the expectation is taken with respect to $P(\theta|D, \xi^{old})$, the variational posterior distribution based on the previous value of ξ. Owing to the EM formulation, each update for ξ corresponds to a monotone improvement to the posterior approximation. Empirically we find that this procedure converges rapidly; only a few iterations are needed. The accuracy of the approximation is considered in the following two sections.

To summarize, the variational approach allows us to obtain a closed form expression for the posterior predictive distribution in logistic regression:

$$P(S|X, \mathcal{D}) = \int P(S|X, \theta) P(\theta|\mathcal{D}) d\theta, \tag{13}$$

where the posterior distribution $P(\theta|\mathcal{D})$ comes from making a single pass through the data set $\mathcal{D} = \{D^1, \ldots, D^T\}$, applying the updates in eq. (10) and eq. (11) after optimizing the associated variational parameters at each step. The predictive lower bound $\underline{P}(S^t|X^t, \mathcal{D})$ takes the form:

$$\log \underline{P}(S^t|X^t, \mathcal{D}) = \log g(\xi_t) - \xi_t/2 + \lambda(\xi_t)\xi_t^2 - \frac{1}{2}\mu^T \Sigma^{-1} \mu + \frac{1}{2}\mu_t^T \Sigma_t^{-1} \mu_t + \frac{1}{2} \log \frac{|\Sigma_t|}{|\Sigma|}, \tag{14}$$

for any complete observation D^t, where μ and Σ signify the parameters in $P(\theta|\mathcal{D})$ and the subscript t refers to the posterior $P(\theta|\mathcal{D}, D^t)$ found by augmenting the data set to include the point D^t.

We note finally that the variational Bayesian calculations presented above need not be carried out sequentially. We could compute a variational approximation to the posterior probability $P(\theta|\mathcal{D})$ by introducing (separate) transformations for each of the logistic functions in

$$P(\mathcal{D}|\theta) = \prod_t P(S^t|X^t, \theta) = \prod_t g\left((2S^t - 1)\, \theta^T X^t\right) \tag{15}$$

The resulting variational parameters would have to be optimized jointly rather than one at a time. We believe the sequential approach provides a cleaner solution.

3 Accuracy of the variational method

The logistic function is shown in Figure 2(a), along with a variational approximation for $\xi = 2$. As we have noted, for each value of the variational parameter ξ, there is a particular point x where the approximation is exact; for the remaining values of x the approximation is a lower bound.

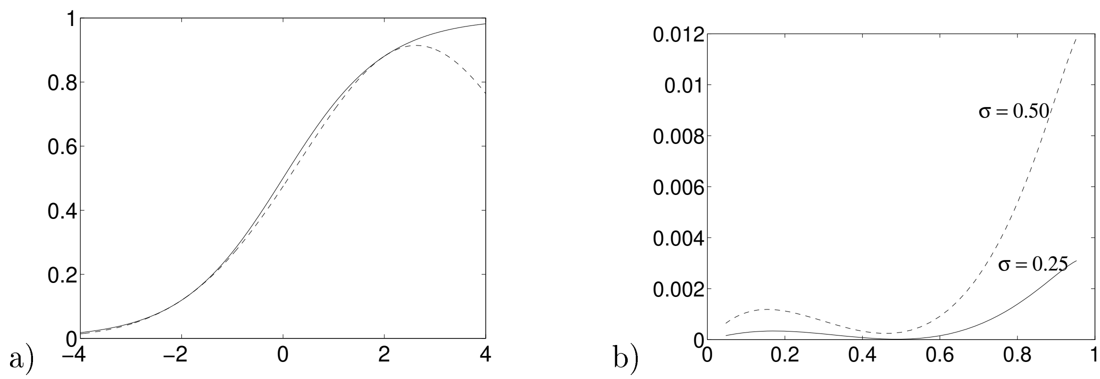


Figure 2: a) The logistic function (solid line) and its variational form (dashed line) for $\xi = 2$. b) The difference between the predictive likelihood and its variational approximation as a function of $g(\mu')$, as described in the text.

Integrating eq. (9) over the parameters we obtain a lower bound on the predictive probability of an observation. The tightness of this lower bound is a measure of accuracy of the approximation. To assess the variational approximation according to this measure, we compared the lower bound to the true predictive likelihood that was evaluated numerically. Note that for a single observation, the evaluation of the predictive likelihood can be reduced to a one-dimensional integration problem:

$$\int P(S|X,\theta)P(\theta)d\theta = \int g((2S-1)\theta^T X)P(\theta)d\theta = \int_{-\infty}^{\infty} g(\theta')P'(\theta')d\theta' \quad (16)$$

where the effective prior $P'(\theta')$ is a Gaussian with mean $\mu' = (2S-1)\mu^T X$ and variance $\sigma^2 = X^T \Sigma X$ where the actual prior distribution $P(\theta)$ has mean μ and covariance Σ. This reduction has no effect on the accuracy of the variational transformation and thus it can be used in evaluating the overall accuracy. Figure 2(b) shows the difference between the true predictive probability and the variational lower bound for various settings of the effective mean μ' and variance σ^2, with ξ optimized separately for each different values of μ' and σ^2. The fact that the variational approximation is a lower bound means that the difference in the predictive likelihood is always positive.

We emphasize that the tightness of the lower bound is not the only relevant measure of accuracy. Indeed, while a tight lower bound on the predictive probability assures us that the associated posterior distribution is highly accurate, the converse is not true in general. In other words, a poor lower bound does not necessarily imply a poor approximation to the posterior distribution at the point of interest, only that we no longer have any guarantees of good accuracy. In practice, we expect the accuracy of the posterior to be more important than that of the predictive probability since errors in the posterior run the risk of accumulating in the course of the sequential estimation procedure. We defer the evaluation of the posterior accuracy to the following section where comparisons are made to related methods.

4 Comparison to other methods

There are other sequential approximation methods that yield closed form posterior parameter distributions in logistic regression models. The method most closely related to ours is that of Spiegelhalter and Lauritzen (1990), which we refer to as the S-L approximation in this paper. Their method is based on the Laplace approximation; that is, they utilize a local quadratic approximation to the complete log-likelihood centered at the prior mean μ. The parameter updates that implement this approximation are similar in spirit to the variational updates of eq. (10) and eq. (11):

$$\Sigma_{post}^{-1} = \Sigma^{-1} + \hat{p}(1-\hat{p})\, XX^T \tag{17}$$

$$\mu_{post} = \mu + (S-\hat{p})\Sigma_{post}\, X \tag{18}$$

where $\hat{p} = g(\mu^T X)$. Since there are no additional adjustable parameters in this approximation, it is simpler than the variational method; however, we would expect this lack of flexibility to translate into less accurate posterior estimates.

We compared the accuracy of the posterior estimates for the two methods in the context of a single observation. To simplify the comparison we utilized the reduction described in the previous section. Since the accuracy of neither method is affected by this reduction, it suffices for our purposes here to carry out the comparison in this simpler setting.[1] The posterior probability of interest was therefore $P(\theta'|D) \propto g(\theta')P(\theta')$, computed for various choices of values for the prior mean μ' and the prior standard deviation σ. The correct posterior mean and standard deviations were obtained numerically. Figures 3 and 4 present the results. We plot signed differences in comparing the obtained posterior means to the correct ones; relative errors were used for the posterior standard deviations. The error measures were left signed to reveal any systematic biases. Note that the posterior mean from the variational method is not guaranteed to be a lower bound on the true mean. Such guarantees can be given only for the predictive likelihood. As can be seen in Figures 3(a) and 4(a) the variational method yields significantly more accurate estimates of the posterior means, for both values of the prior variance. For the posterior variance, the S-L estimate and the variational estimate appear to yield roughly comparable accuracy for the small value of the prior variance (Figure 3(b)); however, for the larger prior variance, the variational approximation is superior (Figure 4(b)). We note that the variational method consistently underestimates the true posterior variance; a fact that could be used to refine the approximation. Finally, in terms of the KL-divergences between the approximate and true posteriors, the variational method and the S-L approximation are roughly equivalent for the small prior variance; and again the variational method is superior for the larger value of the prior variance. This is shown in Figure 5.

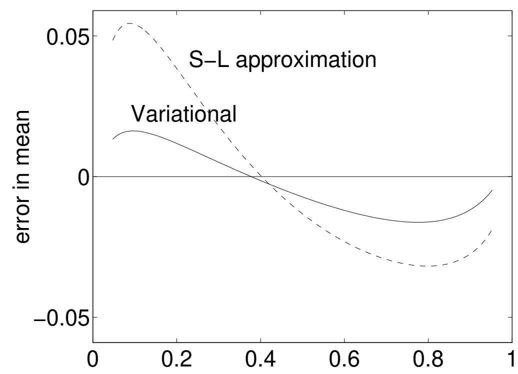
5 Extension to belief networks

A belief network is a probabilistic model over a set of variables $\{S_i\}$ that are identified with the nodes in an acyclic directed graph. Letting $\pi(i)$ denote the set of parents of node S_i in the graph, we define the joint distribution associated with the belief network as the following product:

$$P(S_1, \ldots, S_n) = \prod_i P(S_i|S_{\pi(i)}). \tag{19}$$

We refer to the conditional probabilities $P(S_i|S_{\pi(i)})$ as the "local probabilities" associated with the belief network.

[1]Note that the true posterior distribution over θ can be always recovered from the posterior computed for the one-dimensional reduced parameter $\theta' = \theta^T X$.

a) 

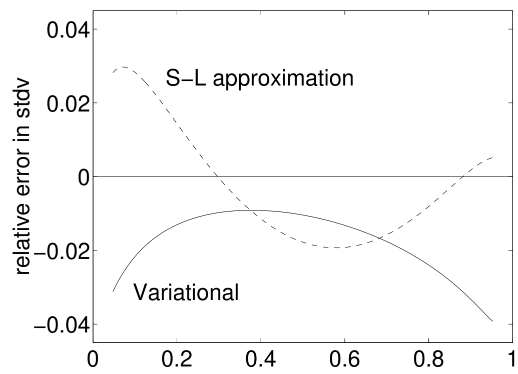
b) 

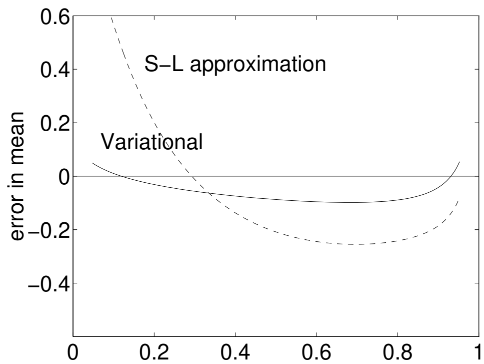
Figure 3: a) The errors in the posterior means as a function of $g(\mu')$, where μ' is the prior mean. Here $\sigma = 1$ for the prior. b) The relative errors in the posterior standard deviations as a function of $g(\mu')$. Again $\sigma = 1$ for the prior distribution.

In this section we extend our earlier work in this paper and consider belief networks in which logistic regression is used to define the local probabilities (such models have been studied in a non-Bayesian setting by Neal 1992 and by Saul, Jaakkola, & Jordan 1994). Thus we introduce parameter vectors θ_i, one for each binary variable S_i, and consider models in which each local probability $P(S_i|S_{\pi(i)}, \theta_i)$ is a logistic regression of node S_i on its parents $S_{\pi(i)}$.

To simplify the arguments in the following sections, we will consider augmented belief networks in which the parameters themselves are treated as nodes in the belief network (see Figure 6). This is a standard device in the belief network literature and is of course natural within the Bayesian formalism.

5.1 Complete cases

A "complete case" refers to a data point in which all of the variables $\{S_i\}$ are observed. If all of the data points are complete cases, then the methods that we developed in the previous section apply immediately to belief networks. This can be seen as follows. Consider the Markov blankets associated with each of the parameters (Figure 6(a)). For complete cases each of the nodes within the Markov blanket for each of the parameters is observed (shaded

a) 

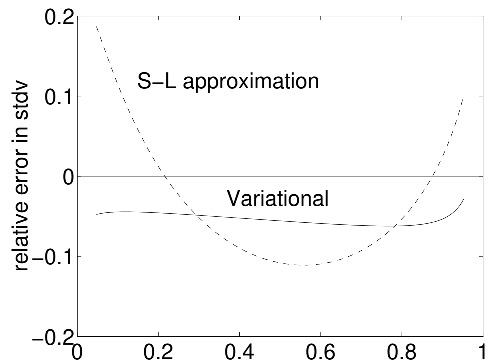
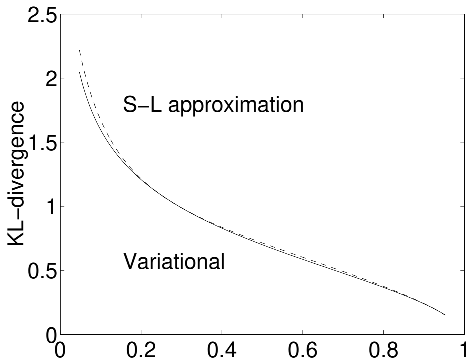
b) 

Figure 4: The plots are the same as in Figure 3, but now $\sigma = 2$ for the prior distribution.

in the diagram). By the independence semantics of belief networks, this implies that the posterior distributions for the parameters are independent of one another (conditioned on the observed data). Thus the problem of estimating the posterior distributions for the parameters reduces to a set of n independent subproblems, each of which is a Bayesian logistic regression problem. We apply the methods developed in the previous sections directly.

5.2 Incomplete cases

The situation is substantially more complex when there are incomplete cases in the data set. Incomplete cases imply that we no longer have all the Markov blankets for the parameters in the network. Thus dependencies can arise between the parameter distributions in different conditional models. Let us consider this situation in some detail. A missing value implies that the observations arise from a marginal distribution obtained by summing over the missing values of the unobserved variables. The marginal distribution is thus a mixture distribution, where each mixture component corresponds to a particular configuration of the missing variables. The weight assigned to that component is essentially the posterior probability of the associated configuration (Spiegelhalter & Lauritzen 1990). Note that the dependencies arising from the missing values in the observations can make the network quite densely connected (a missing value for a node effectively connects all of the neighboring nodes in the graph). The dense connectivity leaves little

a) 

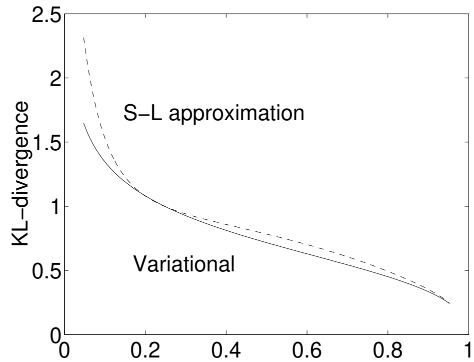
b) 

Figure 5: KL-divergences between the approximate and the true posterior distribution as a function of $g(\mu')$. a) $\sigma = 2$ for the prior. b) $\sigma = 3$. The two approximation methods have (visually) identical curves for $\sigma = 1$.

structure to be exploited in the exact probabilistic computations in these networks and tends to make exact probabilistic calculations intractable.

Our approach to developing Bayesian methods for belief networks with missing variables combines two variational techniques. In particular, we augment the ξ-transformation introduced earlier with a second variational transformation that we refer to as a *q-transformation*. While the purpose of the ξ-transformation is to convert a local conditional probability into a form that can be integrated analytically, the purpose of the q-transformation is to approximate the effect of marginalizing across missing values associated with one or more parents.[2] Intuitively, the q-transformation "fills in" the missing values, allowing the variational transformation for complete data to be invoked. The overall result is a closed-form approximation to the marginal posterior.

The correct marginalization across missing variables is a global operation that affects all of the conditional models that depend on the variables being marginalized over. Under the variational approximation that we describe below, marginalization is a local operation that acts individually on the relevant conditional models.

[2]Treating the parameter as a parent node helps to emphasize the similarity between these two variational transformations. The principal difference is that a parameter node has only a single child, while in general parents have multiple children.

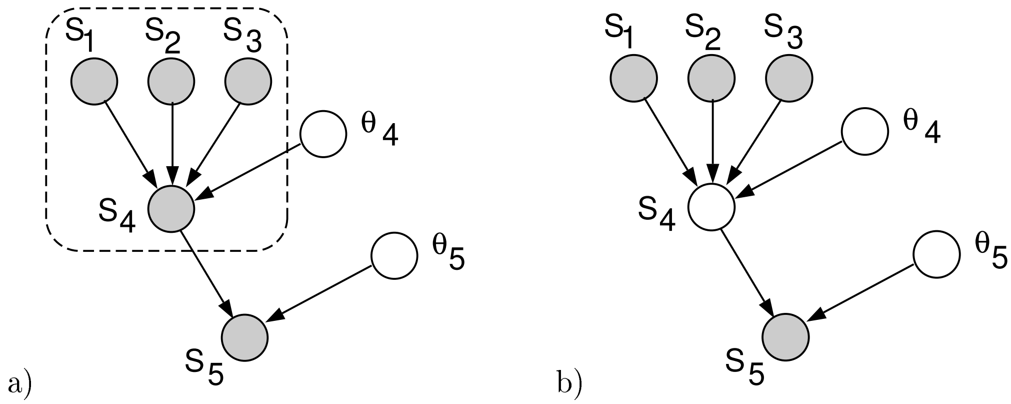


Figure 6: a) A complete observation (shaded variables) and the Markov blanket (dashed line) associated with the parameters θ_4. b) An observation where the value of S_4 is missing (unshaded in the figure).

5.2.1 Approximate marginalization

Consider the problem of marginalizing over a set of variables S' under a joint distribution:

$$P(S_1, \ldots, S_n|\theta) = \prod_i P(S_i|S_{\pi(i)}, \theta_i). \tag{20}$$

If we performed the marginalization exactly, then the resulting distribution would not retain the same factorization as the original joint (assuming S' is involved in more than one of the conditionals); this can be seen from:

$$\sum_{S'} \prod_i P(S_i|S_{\pi(i)}, \theta_i) = \left[\prod_{i''} P(S_{i''}|S_{\pi(i'')}, \theta_{i''})\right] \sum_{S'} \prod_{i'} P(S_{i'}|S_{\pi(i')}, \theta_{i'}), \tag{21}$$

where we have partitioned the product into the set of factors that depend on S' (indexed by i') and those that do not (indexed by i''). Marginalization is not generally a local operation on the individual node probabilities $P(S_i|S_{\pi(S_i)}, \theta_i)$. Maintaining such locality, a desirable goal for computational reasons, can be achieved if we forgo exact marginalization and instead consider approximations. In particular, we describe a variational approximation that preserves locality at the expense of providing a lower bound on the marginal probability instead of an exact result.

To obtain the desired variational transformation, we again exploit a convexity property. In particular, for a given sequence $p_i,\ i \in \{1, \ldots, n\}$, consider the geometric average $\prod_i p_i^{q_i}$, where q_i is a probability distribution. It is well known that the geometric average is less than or equal to the arithmetic average $\sum_i q_i p_i$. (This can be easily established via an invocation of Jensen's inequality). We can exploit this fact as follows. Consider an arbitrary distribution $q(S')$, and rewrite the marginalization operation in the following way:

$$\sum_{S'} P(S_1, \ldots, S_n|\theta) = \sum_{S'} q(S') \left[\frac{P(S_1, \ldots, S_n|\theta)}{q(S')} \right] \tag{22}$$

$$\geq \prod_{S'} \left[\frac{P(S_1, \ldots, S_n|\theta)}{q(S')} \right]^{q(S')} \tag{23}$$

$$= C(q) \prod_i \left[\prod_{S'} P(S_i|S_{\pi(i)}, \theta_i)^{q(S')} \right], \tag{24}$$

where the inequality comes from transforming the average over the bracketed term (with respect to the distribution q) into a geometric average. The third line follows from plugging in the form of the joint distribution and exchanging the order of the products. The logarithm of the multiplicative constant $C(q)$ is the entropy of the variational distribution q:

$$C(q) = \prod_{S'} \left[\frac{1}{q(S')} \right]^{q(S')} \quad \text{and therefore} \quad \log C(q) = -\sum_{S'} q(S') \log q(S') \tag{25}$$

Let us now make a few observations about the result in Eq. (24). First, note that the lower bound in this equation has the same factored form as the original joint probability. In particular, we define the *q-transformation* of the ith local conditional probability as follows:

$$\tilde{P}(S_i|S_{\pi(i)}, \theta_i, q) = \prod_{S'} P(S_i|S_{\pi(i)}, \theta_i)^{q(S')}; \tag{26}$$

the lower bound in Eq. (24) is then a product of these q-transformations. Second, note that all the conditionals are transformed by the same distribution q. A change in q can thus affect all the transformed conditionals. This means that the dependencies between variables S that would have resulted from exact marginalization over S' have been replaced with "effective dependencies" through a shared variational distribution q.

While the bound in Eq. (24) holds for an arbitrary variational distribution $q(S')$, to obtain a tight bound we need to optimize across $q(S')$. In practice this involves choosing a constrained class of distributions and optimizing across the class. The simplest form of variational distribution is the completely factorized distribution:

$$q(S') = \prod_{i=1}^{m} q_i(S'_i), \tag{27}$$

which yields a variational bound which is traditionally referred to as the "mean field approximation." This simplified approximation is appropriate in dense models with a relatively large number of missing values. More generally, one can consider structured variational distributions involving partial factorizations that correspond to tractable substructures in the graphical model (cf. Saul & Jordan, 1996). We consider this topic further in the following two sections.

Although the main constraint on the choice of $q(S')$ is the computational one associated with evaluation and optimization, there is one additional constraint that must be borne in mind. In particular, the q-transformed conditional probabilities must be in a form such that a subsequent ξ-transformation can be invoked, yielding as a result a tractable Bayesian integral. A simple way to meet this constraint is to require that the variational distribution $q(S')$ should not depend on the parameters θ. As we discuss in the following section, in this case all of the q-transformations simply involve products of logistic functions, which behave well under the ξ-transformation.

5.2.2 Bayesian parameter updates

The derivation presented in the previous section shows that approximate variational marginalization across a set of variables S' can be viewed as a geometric average of the local conditional probabilities:

$$P(S|S_\pi, \theta) \rightarrow \prod_{S'} P(S|S_\pi, \theta)^{q(S')} \tag{28}$$

where $q(S')$ is the variational distribution over the missing values. Note that while the ξ-transformations are carried out separately for each relevant conditional model, the variational distribution q associated with the missing values is the same across all the q-transformations.

Given the transformation in eq. (28), the approximate Bayesian updates are obtained readily. In particular, when conditioning on a data point that has missing components we first apply the q-transformation. This effectively fills in the missing values, resulting in a transformed joint distribution that factorizes as in the case of complete observations. The posterior parameter distributions therefore can be obtained independently for the parameters associated with the transformed local probabilities.

Two issues need to be considered. First, the transformed conditional probabilities (cf. Eq. (28)) are products of logistic functions and therefore more complicated than before. The ξ-transformation method, however, transforms each logistic function into an exponential with quadratic dependence on the parameters. Products of such transforms are also exponential with quadratic dependence on the parameters. Thus the approximate likelihood will again be Gaussian and if the prior is a multivariate Gaussian the approximate posterior will also be Gaussian.

The second issue is the dependence of the posterior parameter distributions on the variational distribution q. Once again we have to optimize the variational parameters (a distribution in this case) to make our bounds as tight as possible; in particular, we set q to the distribution that maximizes our lower bound. This optimization is carried out in conjunction with the optimization of the ξ parameters for the transformations of the logistic functions, which are also lower bounds. As we show in Appendix B.1, the fact that all of our approximations are lower bounds implies that we can again devise an EM algorithm to perform the maximization. The updates that are derived in the Appendix are as follows:

$$\Sigma_{pos_i}^{-1} = \Sigma_i^{-1} + 2\lambda(\xi_i)\, E\left\{ S_{\pi(i)} S_{\pi(i)}^T \right\} \tag{29}$$

$$\mu_{pos_i} = \Sigma_{pos_i} \left[\, \Sigma_i^{-1}\mu_i + E\left\{ (S_i - 1/2) S_{\pi(i)} \right\} \right] \tag{30}$$

where $S_{\pi(i)}$ is the vector of parents of S_i, and the expectations are taken with respect to the variational distribution q.

5.2.3 Numerical evaluation

In this section, we provide a numerical evaluation of our proposed combination of q-transformation and ξ-transformation. We study a simple graph that consists of a single node S and its parents S_π. In contrast to the simple logistic regression case analyzed earlier, the parents S_π are not observed but instead are distributed according to a distribution $P(S_\pi)$. This distribution, which we manipulate directly in our experiments, essentially provides a surrogate for the effects of a pattern of evidence in the ancestral graph associated with node S (cf. Spiegelhalter & Lauritzen 1990).

Our interest is in the posterior probability over the parameters θ associated with the conditional probability $P(S|S_\pi, \theta)$.

Suppose now that we observe $S = 1$. The exact posterior probability over the parameters θ in this case is given by

$$P(\theta|D) \propto \left[\sum_{S_\pi} P(S=1|S_\pi, \theta) P(S_\pi) \right] P(\theta) \tag{31}$$

Our variational method focuses on lower bounding the evidence term in brackets. It is natural to evaluate the overall accuracy of the approximation by evaluating the accuracy of the marginal data likelihood:

$$\begin{aligned} P(D) &= \int \left[\sum_{S_\pi} P(S=1|S_\pi, \theta) P(S_\pi) \right] P(\theta) d\theta \qquad (32) \\ &= \int \left[\sum_{S_\pi} \sigma(\,\theta^T S_\pi\,) P(S_\pi) \right] P(\theta) d\theta. \qquad (33) \end{aligned}$$

We consider two different variational approximations. In the first approximation the variational distribution q is left unconstrained; in the second we use an approximation that factorizes across the parents S_π (the "mean field" approximation). We emphasize that in both cases the variational posterior approximation over the parameters is a single Gaussian.

The results of our experiment are shown in Figures 7 and 8. Each figure displays three curves, corresponding to the exact evaluation of the data likelihood $P(D)$ and the two variational lower bounds. The number of parents in S_π was 5 and the prior distribution $P(\theta)$ was taken to be a zero mean Gaussian with a variable covariance matrix. By the symmetry of the Gaussian distribution and the sigmoid function, the exact value of $P(D)$ was 0.5 in all cases. We considered several choices of $P(S_\pi)$ and $P(\theta)$. In the first case, the $P(S_\pi)$ were assumed to factorize across the parents and for each $S_j \in S_\pi$, $P(S_j = 1) = p$ leaving a single parameter p that specifies the stochasticity of $P(S_\pi)$. A similar setting would arise when applying the mean field approximation in the context of a more general graph. Figure 7 shows the accuracy of the variational lower bounds as a function of p where in Figure 7(a) $P(\theta) = N(0, I/5)$, i.e., the covariance matrix is diagonal with diagonal components set to 1/5, and in Figure 7(b) $P(\theta) = N(0, \Sigma)$ where Σ is a sample covariance matrix of 5 Gaussian random vectors distributed according to $N(0, I/5)$. The results of Figure 7(b) are averaged over 5 independent runs. The choice of scaling in $N(0, I/5)$ is made to insure that $|\sum_j \theta_j| \sim 1$. Both figures indicate that the variational approximations are reasonably accurate and that there is little difference between the two methods.

In Figure 8 we see how the mean field approximation (which is unimodal) deteriorates as the distribution $P(S_\pi)$ changes from a factorized distribution toward a mixture distribution. More specifically, let $P_f(S_\pi|p)$ be the (uniform) factorized distribution discussed above with parameter p and let $P_m(S_\pi)$ be a pure mixture distribution that assigns a probability mass 1/3 to three different (randomly chosen) configurations of the parents S_π. We let $P(S_\pi) = (1 - p_m)P_f(S_\pi|p) + p_m P_m(S_\pi)$, where the parameter p_m controls the extent to which $P(S_\pi)$ resembles a (pure) mixture distribution. Figure 8 illustrates the accuracy of the two variational methods as a function of p_m where in Figure 8(a) $p = 0.1$ and in 8(b) $p = 0.3$. As expected, the mean field approximation deteriorates with an increasing p_m whereas our first variational approximation remains accurate.

6 The dual problem

In the logistic regression formulation (eq. (1)), the parameters θ and the explanatory variables X play a dual or symmetric role (cf. Nadal and Parga 1994). In the Bayesian logistic regression setting, the symmetry is broken by associating the same parameter vector θ with multiple occurences of the explanatory variables X as shown in Figure 9. Alternatively, we may break the symmetry by associating a single instance of the explanatory variable X with multiple realizations of θ. In this sense the explanatory variables X play the role of parameters while θ functions as a continuous latent variable. The dual of the Bayesian regression model is thus a latent variable density model over a binary response variable S. Graphically, in the dual interpretation we have a single "parameter" node for X whereas separate nodes are required for different realizations of θ (illustrated as $\theta^{(i)}$ in the figure) to explain successive

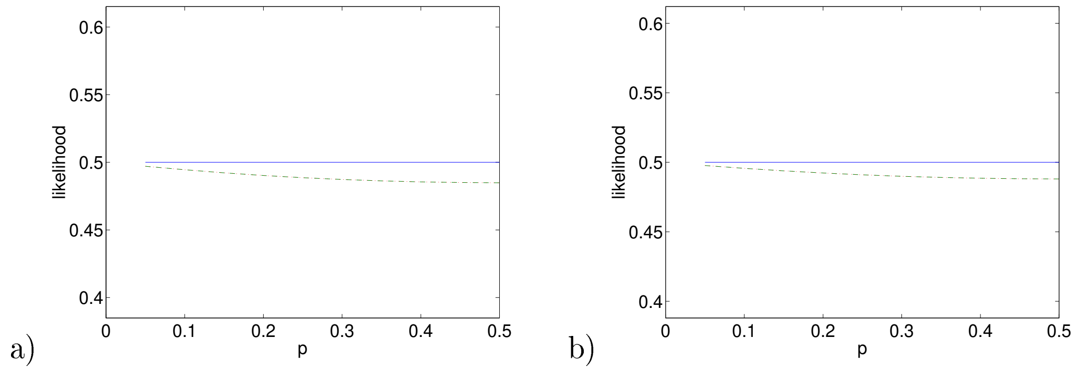


Figure 7: Exact data likelihood (solid line), variational lower bound 1 (dashed line), and variational lower 2 (dotted line) as a function of the stochasticity parameter p of $P(S_\pi)$. In (a) $P(\theta) = N(0, I/5)$ and in (b) $P(\theta) = N(0, \Sigma)$ where Σ is a sample covariance of 5 random vectors distributed according to $N(0, I/5)$.

observations $S^{(i)}$. While a latent variable density model over a single binary variable is not particularly interesting, we can generalize the response variable S to a vector of binary variables $S = [S_1, \ldots, S_n]^T$ where each component S_i has a distinct set of "parameters" $X_i = [X_{i1} \ldots X_{im}]^T$ associated with it. The latent variables θ, however, remain in this dual interpretation the same for all S_i. We note that strictly speaking the dual interpretation would require us to assign a prior distribution over the new "parameters" vectors X_i. For simplicity, however, we omit this consideration and treat X_i simply as adjustable parameters. The resulting latent variable density model over binary vectors is akin to the standard factor analysis model (see e.g. Everitt 1984). This model has already been used to facilitate visualization of high dimensional binary vectors (Tipping 1999).

We now turn to a more technical treatment of this latent variable model. The joint distribution is given by

$$P(S_1, \ldots, S_n | X) = \int \left[\prod_i P(S_i | X_i, \theta) \right] P(\theta) d\theta \qquad (34)$$

where the conditional probabilities for the binary observables are logistic regression models

$$P(S_i | X_i, \theta) = g\left((2S_i - 1) \textstyle\sum_j X_{ij} \theta_j \right) \qquad (35)$$

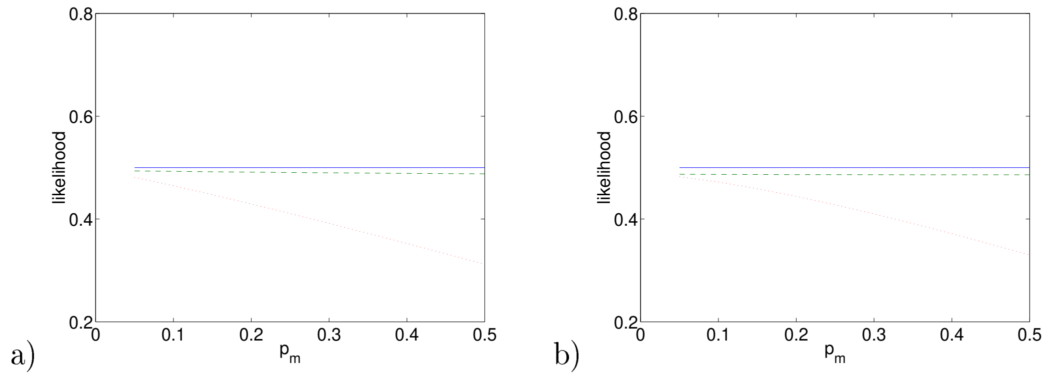


Figure 8: Exact data likelihood (solid line) and the two variational lower bounds (dashed and dotted lines respectively) as a function of the mixture parameter p_m. In (a) $p = 0.1$ and in (b) $p = 0.3$.

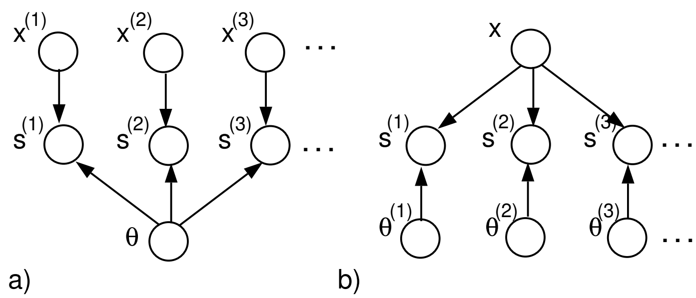


Figure 9: a) Bayesian regression problem. b) The dual problem.

We would like to use the EM- algorithm for parameter estimation. To achieve this we again exploit the variational transformations. The transformations can be introduced for each of the conditional probability in the joint distribution and optimized separately for every observation $D^t = \{S_1^t, \ldots, S_n^t\}$ in the database consisting only of the values of the binary output variables. As in the logistic regression case, the transformations change the unwieldy conditional models into simpler ones that depend on the parameters only quadratically in the exponent. The variational evidence, which is a product of the transformed conditional probabilities, retains the same property. Consequently, under the variational approximation, we can compute the posterior distribution over the latent variables θ in closed form. The mean and the covariance of this posterior can be obtained analogously to the regression case giving

$$\Sigma_t^{-1} = \Sigma^{-1} + \sum_i 2\lambda(\xi_i^t)\, X_i X_i^T \tag{36}$$

$$\mu_t = \Sigma_t \left[\Sigma^{-1}\mu + \sum_i (S_i^t - 1/2) X_i \right] \tag{37}$$

The variational parameters ξ_i^t associated with each observation and the conditional model can be updated using eq. (12) where X is replaced with X_i, now the vector of parameters associated with the i^{th} conditional model.

We can solve the M-step of the EM-algorithm by accumulating sufficient statistics for the parameters X_i, μ, Σ based on the closed form posterior distributions corresponding to the observations in the data set. Omitting the algebra, we obtain the following explicit updates for the parameters:

$$\Sigma \leftarrow \frac{1}{T} \sum_t \Sigma_t \tag{38}$$

$$\mu \leftarrow \frac{1}{T} \sum_t \mu_t \tag{39}$$

$$X_i \leftarrow A_i^{-1} b_i \tag{40}$$

where

$$A_i = \sum_t 2\lambda(\xi_i^t)\, (\Sigma_t + \mu_t \mu_t^T) \tag{41}$$

$$b_i = \sum_t (S_i^t - 1/2) \mu_t \tag{42}$$

and the subscript t denotes the quantities pertaining to the observation $\mathcal{D}_t$. Note that since the variational transformations that we expoited to arrive at these updates are all lower bounds, the M-step necessarily results in a monotonically increasing *lower bound* on the log-probability of the observations. This desirable monotonicity property is unlikely to arise with other types of approximation methods, such as the Laplace approximation.

7 Discussion

We have exemplified the use of variational techniques in the setting of Bayesian parameter estimation. We found that variational methods can be exploited to yield closed form expressions that approximate the posterior distributions for the parameters in logistic regression. The methods apply immediately to a Bayesian treatment of logistic belief networks with complete data. We also showed how to combine mean field theory with our variational transformation and thereby treat belief networks with missing data. Finally, our variational techniques lead to an exactly solvable EM algorithm for a latent variable density model—the dual of the logistic regression problem.

It is also of interest to note that our variational method provides an alternative to the standard iterative Newton-Raphson method for maximum likelihood estimation in logistic regression (an algorithm known as "iterative reweighted least squares" or "IRLS"). The advantage of the variational approach is that it guarantees monotone improvement in likelihood. We present the derivation of this algorithm in Appendix C.

Finally, for an alternative perspective on the application of variational methods to Bayesian inference, see Hinton and van Camp (1993) and MacKay (1997). These authors have developed a variational method known as "ensemble learning," which can be viewed as a mean field approximation to the marginal likelihood.

References

J. Bernardo and A. Smith (1994). *Bayesian Theory.* New York: Wiley.

D. Heckerman, D. Geiger, and D. Chickering (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* **20**, 197-244.

B. Everitt (1984). *An Introduction to Latent Variable Models.* Cambridge University Press.

A. Gelman (1995). *Bayesian Data Analysis.* Boca Raton, FL: CRC Press.

W. Gilks, S. Richardson and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

G. Hinton and D. van Camp (1993). Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the 6th*

Annual Workshop on Computational Learning Theory. New York: ACM Press.

M. Jordan, Z. Ghahramani, T. Jaakkola and L. Saul (1999). An introduction to variational methods in graphical models. In M. I. Jordan, (Ed.), *Learning in Graphical Models*, Cambridge, MA: MIT Press.

P. McCullagh & J. Nelder (1983). *Generalized Linear Models.* London: Chapman and Hall.

D. MacKay (1997). Ensemble learning for hidden Markov models. Unpublished manuscript. Department of Physics, University of Cambridge. Available on the web at `http://wol.ra.phy.cam.ac.uk/mackay`.

J-P. Nadal and N. Parga (1994). Duality between learning machines: A bridge between supervised and unsupervised learning. *Neural Computation* **6**(3):491-508

R. Neal (1992). Connectionist learning of belief networks. *Artificial Intelligence* **56**: 71-113.

R. Neal (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical report CRG-TR-93-1, University of Toronto.

G. Parisi (1988). *Statistical field theory.* Redwood City, CA: Addison-Wesley.

R. Rockafellar (1972). *Convex Analysis.* Princeton University Press.

J. Rustagi (1976). *Variational Methods in Statistics.* New York: Academic Press.

L. Saul, & M. Jordan (1996). Exploiting tractable substructures in intractable networks. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8.* Cambridge MA: MIT Press.

L. Saul, T. Jaakkola, and M. Jordan (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research* **4**: 61-76.

D. Spiegelhalter and S. Lauritzen (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20**: 579-605.

M. Tipping (1999). Probabilistic visualisation of high-dimensional binary data. To appear in *Advances in Neural Information Processing Systems 11.*

A. Thomas, D. Spiegelhalter and W. Gilks (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In *Bayesian Statistics 4*. Clarendon Press.

A Optimization of the variational parameters

To optimize the variational approximation of eq. (9) in the context of an observation $D = \{S, X_1, \ldots, X_r\}$ we formulate an EM algorithm to maximize the predictive likelihood of this observation with respect to ξ. In other words, we find ξ that maximizes the right hand side of

$$\int P(S|X,\theta)P(\theta)d\theta \geq \int \underline{P}(S|X,\theta,\xi)P(\theta)d\theta \tag{43}$$

In the EM formalism this is achieved by iteratively maximizing the expected complete log-likelihood given by

$$Q(\xi|\xi^{old}) = E\left\{\log \underline{P}(S|X,\theta,\xi)P(\theta)\right\} \tag{44}$$

where the expectation is over $P(\theta|D,\xi^{old})$. Taking the derivative of Q with respect to ξ and setting it to zero leads to

$$\frac{\partial}{\partial\xi}Q(\xi|\xi^{old}) = -\frac{\partial\lambda(\xi)}{\partial\xi}\left[E\,(\theta^T X)^2 - \xi^2\right] = 0 \tag{45}$$

As $\lambda(\xi)$ is a monotonically decreasing function[3] the maximum is obtained at

$$\xi^2 = E\,(\theta^T X)^2 \tag{46}$$

By substituting ξ for ξ^{old} above, the procedure can be repeated. Each such iteration yields a better approximation in the sense of eq. (43).

[3]This holds for $\xi \geq 0$. However, since $\underline{P}(S|X,\theta,\xi)$ is a symmetric function of ξ, assuming $\xi \geq 0$ has no effect on the quality of the approximation.

B Parameter posteriors

To fill-in the possible missing values in an observation $D_t = \{S_i^t, \ldots, S_n^t\}$ we employ the q-transformations described in the text. As a result, the joint distribution after the approximate marginalization factorizes as with complete observations. Thus the posterior distributions for the parameters remain independent across the different conditional models and can be computed separately. Thus

$$P(\theta_i | D_t, q) \propto \left[\prod_{S'} P(S_i | S_{\pi(i)}, \theta_i)^{q(S')} \right] P(\theta_i) \tag{47}$$

The form of this posterior, however, remains at least as unwieldy as the Bayesian logistic regression problem considered earlier in the paper. Proceeding analogously, we transform the logistic functions as in Eq. (7) corresponding to each of the conditional probabilities in the product and obtain

$$\begin{aligned}
P(\theta_i | D_t, q, \xi_i) &\propto \left[\prod_{S'} P(S_i | S_{\pi(i)}, \theta_i, \xi_i)^{q(S')} \right] P(\theta_i) && (48)\\
&= \left[\prod_{S'} \left\{ g(\xi_i)\, e^{(H_{S_i} - \xi_i)/2 - \lambda(\xi_i)(H_{S_i}^2 - \xi_i^2)} \right\}^{q(S')} \right] P(\theta_i) && (49)\\
&= \left[g(\xi_i)\, e^{\sum_{S'} q(S') \left\{ (H_{S_i} - \xi_i)/2 - \lambda(\xi_i)(H_{S_i}^2 - \xi_i^2) \right\}} \right] P(\theta_i) && (50)\\
&= \left[g(\xi_i)\, e^{(E\{H_{S_i}\} - \xi_i)/2 - \lambda(\xi_i)(E\{H_{S_i}^2\} - \xi_i^2)} \right] P(\theta_i) && (51)\\
&\equiv \underline{P}(S_i | S_{\pi(i)}, \theta_i, \xi_i, q) P(\theta_i) && (52)
\end{aligned}$$

where $H_{S_i} = (2S_i - 1)\theta_i^T S_{\pi(i)}$, $S_{\pi(i)}$ is the vector of parents of S_i, and the expectations are with respect to the variational distribution q. For simplicity, we have not let the variational parameter ξ_i vary independently with the configurations of the missing values but assumed it to be the same for all such configurations. This choice is naturally suboptimal but is made here primarily for notational simplicity (the choice may also be necessary in cases where the number of missing values is large). Now, since H_{S_i} is linear in the parameters θ_i, the exponent in Eq. (51) consisting of averages over H_{S_i} and its square with respect to the variational distribution q, stays at most quadratic in the parameters θ_i. A multivariate Gaussian prior will be conjugate to this likelihood, and therefore the posterior will also be Gaussian. The mean μ_{pos_i} and covariance Σ_{pos_i} of such posterior are given by (we omit the algebra)

$$\Sigma_{pos_i}^{-1} = \Sigma_i^{-1} + 2\lambda(\xi_i)\, E\left\{ S_{\pi(i)} S_{\pi(i)}^T \right\} \tag{53}$$

$$\mu_{pos_i} = \Sigma_{pos_i} \left[\, \Sigma_i^{-1}\mu_i + E\left\{ (S_i - 1/2) S_{\pi(i)} \right\} \,\right] \tag{54}$$

Note that this posterior depends both on the distribution q and the parameters ξ. The optimization of these parameters is shown in Appendix B.1.

B.1 Optimization of the variational parameters

We have introduced two variational "parameters": the distribution q over the missing values, and the ξ parameters corresponding to the logistic or ξ-transformations. The metric for optimizing the parameters comes from the fact that the transformations associated with these parameters introduce a lower bound on the probability of the observations. Thus by maximizing this lower bound we find the parameter values that yield the most accurate approximations. We therefore attempt to maximize the right hand side of

$$\begin{aligned} \log P(D_t) &\geq \log \underline{P}(D_t|\xi, q) && (55)\\ &= \log \int \underline{P}(D_t|\theta, \xi, q) P(\theta) d\theta && (56)\\ &= \log \prod_i \int \underline{P}(S_i|S_{\pi(i)}, \theta_i, \xi_i, q) P(\theta_i) d\theta_i + \log C(q) && (57) \end{aligned}$$

where D_t contains the observed settings of the variables. We have used the fact that the joint distribution under our approximations factorizes as with complete cases. Similarly to the case of the simple Bayesian logistic regression considered previously (see Appendix A), we can devise an EM-algorithm to maximize the variational lower bound with respect to the parameters q and ξ; the parameters θ can be considered as latent variables in this formulation. The E-step of the EM-algorithm, i.e., finding the posterior distribution over the latent variables, has already been described in Appendix B. Here we will consider in detail only the M-step. For simplicity, we solve the M-step in two phases: the first where the variational distribution is kept fixed and the maximization is over ξ, and the second where these roles are reversed. We start with the first phase.

As the variational joint distribution factorizes, the problem of finding the optimal ξ parameters separates into independent problems concerning each of the transformed conditionals. Thus the optimization becomes analogous to the simple Bayesian logistic regression considered earlier. Two differences exist: first, the posterior over each θ_i is now obtained from Eq. (52); second, we have an additional expectation with respect to the variational distribution q. With these differences the optimization is analogous to the one presented in Appendix A above and we won't repeat it here.

The latter part of our two-stage M-step is new, however, and we will consider it in detail. The objective is to optimize q while keeping the ξ parameters fixed to their previously obtained values. Similarly to the ξ case we construct an EM-algorithm to perform this inner loop optimization:

$$
\begin{aligned}
Q(q|q^{old}) &= E_\theta\{\log \underline{P}(D_t, \theta|\xi, q)\} && (58)\\
&= \sum_i E_{\theta_i}\left\{\log \underline{P}(S_i|S_{\pi(i)}, \theta_i, \xi_i, q) P(\theta_i)\right\} + \log C(q) && (59)
\end{aligned}
$$

where the first expectation is with respect to $P(\theta|\xi, q^{old})$, which factorizes across the conditional probabilities as explained previously; the expectations E_{θ_i} are over the component distributions $P(\theta_i|\xi_i, q^{old})$, obtained directly from Eq. (52). Let us now insert the form of the transformed conditional probabilities, $P(S_i|S_{\pi(i)}, \theta_i, \xi_i, q)$, into the above definition of the Q function. For clarity we will omit all the terms with no dependence on the variational distribution q. We obtain:

$$
\begin{aligned}
Q(q|q^{old}) &= \sum_i E_{\theta_i}\left\{ E_q\{H_{S_i}\}/2 - \lambda(\xi_i)\, E_q\{H_{S_i}^2\} \right\} + \log C(q) + \ldots && (60)\\
&= E_q \sum_i \left\{ E_{\theta_i}\{H_{S_i}\}/2 - \lambda(\xi_i)\, E_{\theta_i}\{H_{S_i}^2\} \right\} + \mathcal{H}(q) + \ldots && (61)
\end{aligned}
$$

where E_q refers to the expectation with respect to the variational distribution q. The second equation follows by exchanging the order of the (mutually independent) expectations E_{θ_i} and E_q. We have also used the fact that $\log C(q)$ is the entropy $\mathcal{H}(q)$ of q (see the text). Recall the notation $H_{S_i} = (2S_i - 1)\theta_i^T S_{\pi(i)}$, where $S_{\pi(i)}$ is a binary vector of parents of S_i. Before proceeding to maximize the Q function with respect to q, we explicate the averages E_{θ_i} in the above formula:

$$
\begin{aligned}
E_{\theta_i}\{H_{S_i}\} &= (2S_i - 1)\mu_{p_i}^T S_{\pi(i)} && (62)\\
E_{\theta_i}\{H_{S_i}^2\} &= \left(\mu_{p_i}^T S_{\pi(i)}\right)^2 + S_{\pi(i)}^T \Sigma_{p_i} S_{\pi(i)} && (63)
\end{aligned}
$$

Here μ_{p_i} and Σ_{p_i} are the mean and the covariance, respectively, of the posterior $P(\theta_i|\xi_i, q^{old})$ associated with the i^{th} conditional model. Simply inserting these back into the expression for the Q function we get

$$
\begin{aligned}
Q(q|q^{old}) \;\; = \;\; & E_q \sum_i \Big\{ \, (S_i - 1/2)\mu_{p_i}^T S_{\pi(i)} - \lambda(\xi_i) \, \left(\mu_{p_i}^T S_{\pi(i)}\right)^2 \\
& \qquad -\lambda(\xi_i) \, S_{\pi(i)}^T \Sigma_{p_i} S_{\pi(i)} \, \Big\} + \mathcal{H}(q) + \ldots \qquad (64)
\end{aligned}
$$

Now, some of the binary variables S_i have a value assignment based on the observation D_t and the remaining variables will be averaged over the variational distribution q. Assuming no a priori constraints on the form of the q distribution, the maximizing q is the Boltzmann distribution (see e.g. Parisi 1988):

$$
q(S') \propto \exp\left(\sum_i \Big\{ \, (S_i - 1/2)\mu_{p_i}^T S_{\pi(i)} - \lambda(\xi_i) \, \left(\mu_{p_i}^T S_{\pi(i)}\right)^2 - \lambda(\xi_i) \, S_{\pi(i)}^T \Sigma_{p_i} S_{\pi(i)} \, \Big\} \right) \qquad (65)
$$

Whenever the variational distribution is constrained, however, such as in the case of a completely factorized distribution, we may no longer expect to find the q that maximizes eq. (64). Nevertheless, a locally optimal solution can be found by, for example, sequentially solving

$$
\frac{\partial}{\partial q_k} Q(q|q^{old}) = 0 \qquad (66)
$$

with respect to each of the components $q_k = q_k(S_k = 1)$ in the factorized distribution.

C Technical note: ML estimation

The standard maximum likelihood procedure for estimating the parameters in logistic regression uses an iterative Newton-Raphson method to find the parameter values. While the method is fast, it is not monotonic; i.e., the probability of the observations is not guaranteed to increase after any iteration. We show here how to derive a monotonic, fast estimation procedure for logistic regression by making use of the variational transformation in eq. (7). Let us denote $H_t = (2S_t - 1)\theta^T X_t$ and write the log-probability of the observations as

$$
\begin{aligned}
\mathcal{L}(\theta) &= \sum_t \log P(S_t|X_t,\theta) = \sum_t \log g(H_t) \\
&\geq \sum_t \log g(\xi_t) + (H_t - \xi_t)/2 - \lambda(\xi_t)\left(H_t^2 - \xi_t^2\right) \\
&= \mathcal{L}(\theta,\xi) \qquad (67)
\end{aligned}
$$

The variational lower bound is exact whenever $\xi_t = H_t$ for all t. Although the parameters θ cannot be solved easily from $\mathcal{L}(\theta)$, $\mathcal{L}(\theta,\xi)$ allows a closed form solution for any fixed ξ, since the variational log-probability is a quadratic function of θ. The parameters θ that maximize $\mathcal{L}(\theta,\xi)$ are given by $\theta' = A^{-1}b$ where

$$
A = \sum_t 2\lambda(\xi_t)\, H_t {H_t}^T \text{ and } b = \sum_t (S_t - 1/2) H_t \qquad (68)
$$

Successively solving for θ and updating ξ yields the following chain of inequalities:

$$
\mathcal{L}(\theta) = \mathcal{L}(\theta,\xi) \leq \mathcal{L}(\theta',\xi) \leq \mathcal{L}(\theta',\xi') = \mathcal{L}(\theta') \qquad (69)
$$

where the prime signifies an update and we have assumed that $\xi_t = H_t$ initially. The combined update thus leads to a monotonically increasing log-probability. In addition, the closed form θ-updates make this procedure comparable in speed to the standard Newton-Raphson alternative.