Variational inference for continuous sigmoidal Bayesian networks

Brendan J. Frey
Department of Computer Science, University of Toronto
6 King's College Road, Toronto, Canada M5S 1A4

November 11, 1996

Abstract

Latent random variables can be useful for modelling covariance relationships between observed variables. The choice of whether specific latent variables ought to be continuous or discrete is often an arbitrary one. In a previous paper, I presented a "unit" that could adapt to be continuous or binary, as appropriate for the current problem, and showed how a Markov chain Monte Carlo method could be used for inference and parameter estimation in Bayesian networks of these units. In this paper, I develop a variational inference technique in the hope that it will prove to be more computationally efficient than Monte Carlo methods. After presenting promising *inference* results on a toy problem, I discuss why the variational technique does not work well for *parameter estimation* as compared to Monte Carlo.

Introduction

Inference in multiply-connected Bayesian networks with real-valued random variables is a difficult problem. Methods such as probability propagation (Gallager 1963; Pearl 1986; Lauritzen and Spiegelhalter 1988) are exact only for singly-connected networks, and techniques for converting multiply-connected networks to singly-connected networks often lead to overly-complex cluster variables. Real-valued variables introduce another level of difficulty, since it is not obvious in general how to efficiently represent probability "messages" in this case.

Recently, there has been a surge of interest in inference and parameter estimation in Bayesian networks with discrete-valued variables whose conditional distributions are modelled using logistic regression (McCullagh and Nelder 1983). Approximate inference methods for richly-connected Bayesian networks of this sort have been developed, including Markov chain Monte Carlo methods (Neal 1992), Helmholtz machines (Dayan *et. al.* 1995; Hinton *et. al.* 1995), and variational (sometimes called "mean field") techniques (Saul *et. al.* 1996; Jaakkola *et. al.* 1996).

However, some hidden variables, such as translation or scaling in images of shapes, are best represented using real values. A great deal of work has been done on Gaussian random variables that are linked linearly such that the joint distribution over all variables is also Gaussian (Pearl 1988; Shachter and Kenley 1989; Spiegelhalter 1990; Heckerman and Geiger 1995) — see also "factor analysis" (Everitt 1984). Lauritzen and Wermuth (1989) and Lauritzen, Dawid, Larsen, and Leimer (1990) have included discrete random variables within the linear Gaussian framework. Recently, inference in networks of Gaussian random variables that are linked nonlinearly have been explored by Driver and Morrell (1995). All of these approaches employ probability propagation for inference and so are not well-tailored to richly-connected networks. Also, these approaches tend to assume that all the conditional Gaussian distributions represented by the belief network can be easily derived using information elicited from experts.

Tibshirani (1992) and Bishop *et. al.* (1996) consider nonlinear mappings from a *continuous* latent variable space to a higher-dimensional input space. Mackay (1995) has developed "density networks" that can model both continuous and categorical latent spaces using stochasticity at the top-most network layer. Hofmann and Tresp (1996) consider the case of inference and learning in continuous belief networks that may be richly connected. They use mixture models and Parzen windows to implement conditional densities.

In (Frey 1997), I presented a simple, but versatile, real-valued random "unit" that can operate in several different modes ranging from deterministic to binary stochastic to continuous stochastic. This spectrum of behaviors is controlled by only two parameters. In this paper I consider a variational inference technique for networks of these units. The method is similar in flavor to the one proposed by Jaakkola *et. al.* (1996). However, their method is developed in an "extended domain" that is meant to make inference in networks of *binary* variables more tractable. In the present paper, the variables are real-valued.

Continuous sigmoidal Bayesian networks

The continuous sigmoidal unit is shown in figure 1a. Each unit i contains a parameter σ_i^2, and the input m_i to unit i is determined by the outputs of other units, as described later. Output y_i depends on the values of m_i and σ_i^2, according to a conditional distribution. The conditional distribution for the *presigmoid* activity x_i for unit i is

$$P(x_i|m_i, \sigma_i^2) \equiv \frac{1}{\sqrt{2\pi\sigma_i^2}}\, e^{-(x_i-m_i)^2/2\sigma_i^2}, \tag{1}$$

where m_i and σ_i^2 are the mean and variance for unit i. A *postsigmoid* activity, y_i, is obtained by passing the presigmoid activity through a fixed cumulative Gaussian squashing function:

$$y_i \equiv \Phi(x_i) \equiv \int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2} dz. \tag{2}$$

Including the transformation Jacobian, the postsigmoid distribution for unit i is

$$P(y_i|m_i, \sigma_i^2) = \frac{1}{\Phi'(\Phi^{-1}(y_i))\sqrt{2\pi\sigma_i^2}}\, e^{-(\Phi^{-1}(y_i)-m_i)^2/2\sigma_i^2}, \tag{3}$$

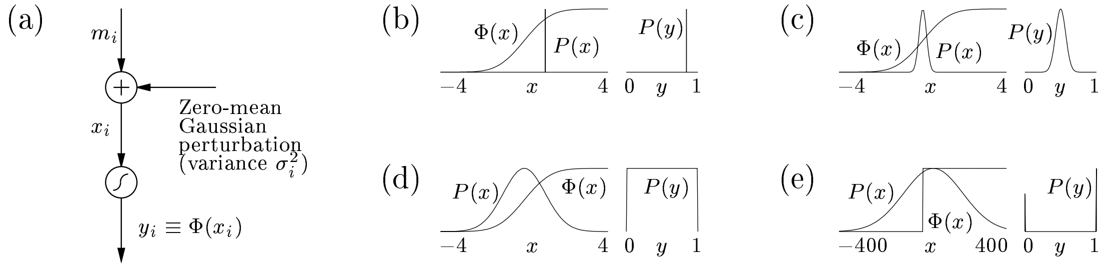


Figure 1: (a) shows the inner workings of the continuous sigmoidal unit. (b) to (e) illustrate four quite different modes of behavior: (b) deterministic mode; (c) stochastic linear mode; (d) stochastic nonlinear mode; and (e) stochastic binary mode (note the different horizontal scale). For the sake of graphical clarity, the density functions are normalized to have equal maxima and the subscripts are left off the variables.

where $\Phi'(x) = \phi(x) \equiv e^{-x^2/2}/\sqrt{2\pi}$. Both $\Phi()$ and $\Phi^{-1}()$ are nonanalytic, so the C-library erf() function is used to implement $\Phi()$ and table lookup with quadratic interpolation is used to implement $\Phi^{-1}()$.

Networks of these units can *represent* a variety of structures. This versatility is brought about by a range of significantly different modes of behavior available to each unit. Figures 1b to 1e illustrate these modes.

Deterministic mode: If the variance of a unit is very small, the postsigmoid activity will be a practically deterministic sigmoidal function of the mean (figure 1a). This mode is useful for representing deterministic nonlinear mappings, such as those found in multi-layer perceptrons.

Stochastic linear mode: For a given mean, if the squashing function is approximately linear over the span of the added noise, the postsigmoid distribution will be approximately Gaussian with the mean and standard deviation linearly transformed (figure 1b). This mode is useful for representing latent Gaussian random variables, such as those used in factor analysis (Everitt 1984).

Stochastic nonlinear mode: If the variance of a unit in the stochastic linear mode is increased so that the squashing function is used in its nonlinear region, a variety of distributions are producible that range from skewed Gaussian to uniform to bimodal (figure 1c)..

Stochastic binary mode: This is an extreme case of the stochastic nonlinear mode. If the variance of a unit is very large, then nearly all of the probability mass will lie near the ends of the interval $(0, 1)$ (figure 1e). For example, for a standard deviation of 150, less than 1% of the mass lies in $(0.1, 0.9)$. In this mode, the postsigmoid activity of unit i appears to be binary with probability of being "on" (*i.e.*, $y_i > 0.5$ or, equivalently, $x_i > 0$):

$$P(i \text{ on}|m_i, \sigma_i^2) = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-(x-m_i)^2/2\sigma_i^2} dx = \int_{-\infty}^{m_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-x^2/2\sigma_i^2} dx = \Phi\left(\frac{m_i}{\sigma_i}\right). \tag{4}$$

In this limit, the proposed units becomes identical to the binary stochastic units of Jaakkola *et. al.* (1996).

If the mean of each unit depends on the activities of other units and there are feedback connections, it is difficult to relate the density in equation 3 to a joint distribution over all unit activities, and simulating the model would require a great deal of computational effort. However, when a top-down topology is imposed on the network (making it a directed acyclic graph), the densities given in equations 1 and 3 can be interpreted as conditional distributions corresponding to a joint distribution over all units. The joint distribution can be expressed as

$$P(\{x_i\}_{\forall i}) = \prod_{i=1}^{N} P(x_i|\{x_j\}_{\forall j<i}) \quad \text{or} \quad P(\{y_i\}_{\forall i}) = \prod_{i=1}^{N} P(y_i|\{y_j\}_{\forall j<i}), \tag{5}$$

where N is the number of units. $P(x_i|\{x_j\}_{j<i})$ and $P(y_i|\{y_j\}_{j<i})$ are the presigmoid and postsigmoid densities of unit i conditioned on the activities of units with lower indices. This ordered arrangement is the foundation of Bayesian networks (Pearl 1988). I consider networks where the mean of each unit be determined by a linear combination of the postsigmoid activities of preceding units:

$$m_i = \sum_{\forall j<i} w_{ij} y_j = \sum_{\forall j<i} w_{ij} \Phi(x_j), \tag{6}$$

where $y_0 \equiv 1$ is used to implement a bias. The variance for each unit is independent of unit activities. So, the presigmoid density for unit i conditioned on the values of preceding units is

$$P(x_i|\{x_j\}_{\forall j<i}) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-(x_i - \sum_{\forall j<i} w_{ij}\Phi(x_j))^2/2\sigma_i^2}. \tag{7}$$

A single sample from the joint distribution can be obtained by traversing the network from top to bottom while computing this density for each unit and drawing an activity for each unit.

Variational inference

Given the activities of a set of visible (observed) variables $\{x_i\}_{i\in V}$, inferring the distribution $P(\{x_j\}_{j\in H}|\{x_i\}_{i\in V})$ over the remaining set of hidden (unobserved) variables $\{x_j\}_{j\in H}$, is in general NP-hard (Cooper 1990). Inference is especially difficult when the variables are real-valued. Efficient algorithms such as *probability propagation* (Gallager 1963; Pearl 1986; Lauritzen and Spiegelhalter 1988) are exact only for singly-connected networks. In (Frey 1997), I used a Markov chain Monte Carlo method called "slice sampling" (Neal 1996) to obtain an approximate sample from $P(\{x_j\}_{j\in H}|\{x_i\}_{i\in V})$, which could then be used for inference. In contrast to both the rather unprincipled approach of applying probability propagation to multiply-connected networks, and the computationally intensive stochastic approach of Monte Carlo, variational inference is a nonstochastic technique that directly

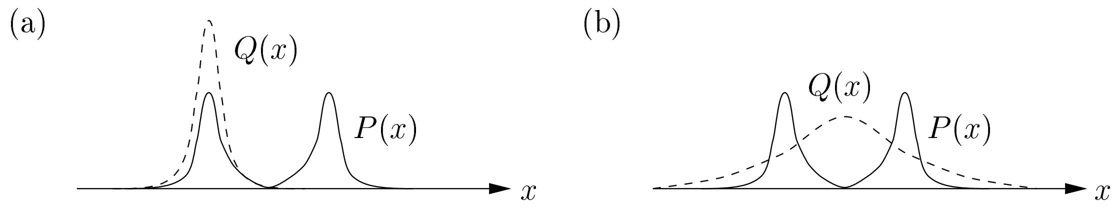


Figure 2: The effect of using (a) $D(Q \parallel P)$ versus (b) $D(P \parallel Q)$ when fitting a variational distribution $Q(x)$ to a univariate distribution $P(x)$.

addresses the quality of inference. Variational inference has recently been championed in the Bayesian network literature by Ghahramani, Jaakkola, Jordan and Saul (Saul *et. al.* 1996; Jaakkola *et. al.* 1996; Ghahramani and Jordan 1996). It is based on a variational interpretation of the expectation maximization algorithm (Neal and Hinton 1993) that was used to devise the Helmholtz machine (Dayan *et. al.* 1995; Hinton *et. al.* 1995).

The idea is to introduce a second parametric distribution $Q(\{x_j\}_{j\in H})$ over the hidden variables, whose parameters are adjusted so as to minimize the "distance" between $Q(\{x_j\}_{j\in H})$ and $P(\{x_j\}_{j\in H}|\{x_i\}_{i\in V})$, given the values of a set of observed variables. Once this optimization is complete, the parametric distribution $Q(\{x_j\}_{j\in H})$ is used as an approximation to $P(\{x_j\}_{j\in H}|\{x_i\}_{i\in V})$. The form of $Q(\{x_j\}_{j\in H})$ and the distance measure are chosen in accordance with the desired properties of the approximation and the tractability of computing the distance and possibly its derivatives (depending on the optimization method used).

Here, I use a simple approximation consisting of a product of Gaussian distributions over the hidden variables,

$$Q(\{x_j\}_{j\in H}) = \prod_{j\in H} Q(x_j), \quad \text{where} \quad Q(x_j) = \frac{1}{\sqrt{2\pi s_j^2}} e^{-(x_j-\mu_j)^2/2s_j^2}, \tag{8}$$

and a relative entropy (Kullback-Liebler) pseudo-distance:

$$D(Q \parallel P) = \int_{\{x_j\}_{j\in H}} Q(\{x_j\}_{j\in H}) \log\Big[\frac{Q(\{x_j\}_{j\in H})}{P(\{x_j\}_{j\in H}|\{x_i\}_{i\in V})}\Big] \prod_{j\in H} dx_j. \tag{9}$$

As shown below, this distance leads to a tractable optimization problem in the μ_j's and s_j's. The choice of using $D(Q \parallel P)$ versus $D(P \parallel Q)$ also depends on our objective. The former places emphasis on not inferring unlikely values of the hidden variables at the cost of not inferring some of the likely values. In contrast, the latter places emphasis on inferring all likely values of the hidden variables at the cost of inferring some of the unlikely values. For example, consider a real-valued univariate distribution $P(x)$ that has two modes, as shown in figure 2. Suppose the variational distribution $Q(x)$ is a Gaussian with a mean and a variance. Figure 2a shows the optimum variational distribution that is obtained by minimizing $D(Q \parallel P)$, whereas figure 2b shows the optimum variational distribution that is obtained by minimizing $D(P \parallel Q)$.

The relative entropy in equation 9 cannot in general be optimized directly because the denominator $P(\{x_j\}_{j\in H}|\{x_i\}_{i\in V})$ in the logarithm cannot in general be expressed in a simple form. A simple form is obtained in the following way. Since $\log P(\{x_i\}_{i\in V})$ does not depend on the μ_j's and s_j's, it can be subtracted from $D(Q \parallel P)$ without changing the optimization. The resulting function $F(Q,P)$ is

$$F(Q,P) = D(Q \parallel P) - \log P(\{x_i\}_{i\in V}) \tag{10}$$

$$= D(Q \parallel P) - \int_{\{x_j\}_{j\in H}} Q(\{x_j\}_{j\in H}) \log P(\{x_i\}_{i\in V}) \textstyle\prod_{j\in H} dx_j$$

$$= \int_{\{x_j\}_{j\in H}} Q(\{x_j\}_{j\in H}) \log\Big[\frac{Q(\{x_j\}_{j\in H})}{P(\{x_j\}_{j\in H}, \{x_i\}_{i\in V})}\Big] \textstyle\prod_{j\in H} dx_j$$

$$= \mathrm{E}\Big[\log \frac{Q(\{x_j\}_{j\in H})}{P(\{x_j\}_{j\in H}, \{x_i\}_{i\in V})}\Big], \tag{11}$$

where $\mathrm{E}[\cdot]$ is an expectation over $Q(\{x_j\}_{j\in H})$. Now, the denominator in the logarithm can be factored using equation 5. Combining this with equations 7 and 8, we get

$$F(Q,P) = \mathrm{E}\Big[\log \frac{\prod_{i\in H} Q(x_i)}{\prod_{i\in H} P(x_i|\{x_j\}_{j<i}) \prod_{i\in V} P(x_i|\{x_j\}_{j<i})}\Big]$$

$$= \sum_{i\in H} \mathrm{E}\Big[\log \frac{Q(x_i)}{P(x_i|\{x_j\}_{j<i})}\Big] + \sum_{i\in V} \mathrm{E}\Big[\log \frac{1}{P(x_i|\{x_j\}_{j<i})}\Big] \tag{12}$$

$$= \sum_{i\in H} \mathrm{E}\Bigg[\log \frac{e^{-(x_i-\mu_i)^2/2s_i^2}/\sqrt{2\pi s_i^2}}{e^{-(x_i-\sum_{\forall j<i} w_{ij}\Phi(x_j))^2/2\sigma_i^2}/\sqrt{2\pi\sigma_i^2}}\Bigg] + \sum_{i\in V} \mathrm{E}\Bigg[\log \frac{1}{e^{-(x_i-\sum_{\forall j<i} w_{ij}\Phi(x_j))^2/2\sigma_i^2}/\sqrt{2\pi\sigma_i^2}}\Bigg]$$

$$= \frac{1}{2}\sum_{i\in H} \mathrm{E}\Big[\log \frac{\sigma_i^2}{s_i^2} - \frac{(x_i-\mu_i)^2}{s_i^2} + \frac{(x_i - \sum_{\forall j<i} w_{ij}\Phi(x_j))^2}{\sigma_i^2}\Big] + \frac{1}{2}\sum_{i\in V} \mathrm{E}\Big[\log 2\pi + \log\sigma_i^2 + \frac{(x_i - \sum_{\forall j<i} w_{ij}\Phi(x_j))^2}{\sigma_i^2}\Big]. \tag{13}$$

After simplification, we get

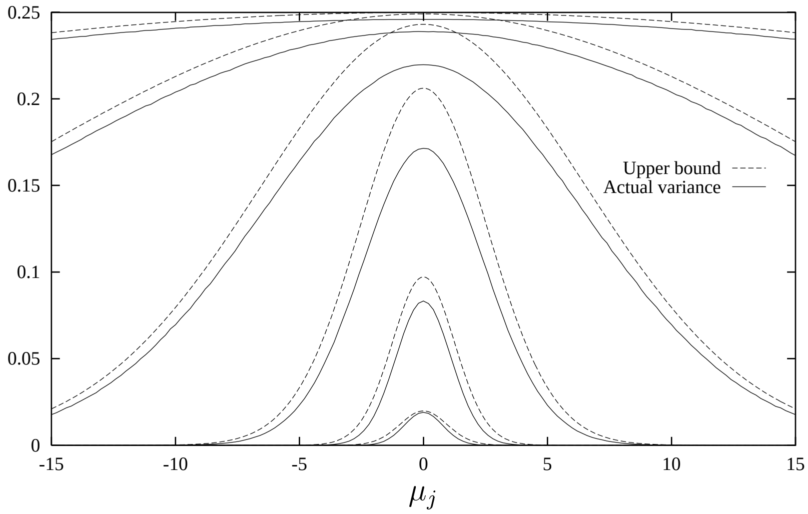


Figure 3: The bound (dashed line) on the postsigmoid variance (solid line).

$$F(Q,P) = \frac{1}{2}\sum_{i\in H}\Big[\log\frac{\sigma_i^2}{s_i^2} + \frac{s_i^2}{\sigma_i^2} - 1 + \frac{1}{\sigma_i^2}\Big[\mu_i - \sum_{\forall j<i} w_{ij}M(\mu_j,s_j)\Big]^2 + \frac{1}{\sigma_i^2}\sum_{\forall j<i} w_{ij}^2 V(\mu_j,s_j)\Big]$$

$$+\frac{1}{2}\sum_{i\in V}\Big[\log 2\pi + \log\sigma_i^2 + \frac{1}{\sigma_i^2}\Big[\mu_i - \sum_{\forall j<i} w_{ij}M(\mu_j,s_j)\Big]^2 + \frac{1}{\sigma_i^2}\sum_{\forall j<i} w_{ij}^2 V(\mu_j,s_j)\Big], \tag{14}$$

where $M(\mu_j, s_j) = \mathrm{E}[\Phi(x_j)]$ and $V(\mu_j, s_j) = \mathrm{E}[(\Phi(x_j) - M(\mu_j, s_j))^2]$ are the postsigmoid mean and variance induced by the Gaussian distribution $Q(x_j)$ over the input x_j to the sigmoid. For observed variables, $M(\mu_j, s_j) = x_j$ and $V(\mu_j, s_j) = 0$. The term $[\mu_i - \sum_{\forall j<i} w_{ij}M(\mu_j, s_j)]^2$ encourages a low discrepancy between the inferred input μ_i to unit i and the "mean field" input $\sum_{\forall j<i} w_{ij}M(\mu_j, s_j)$.

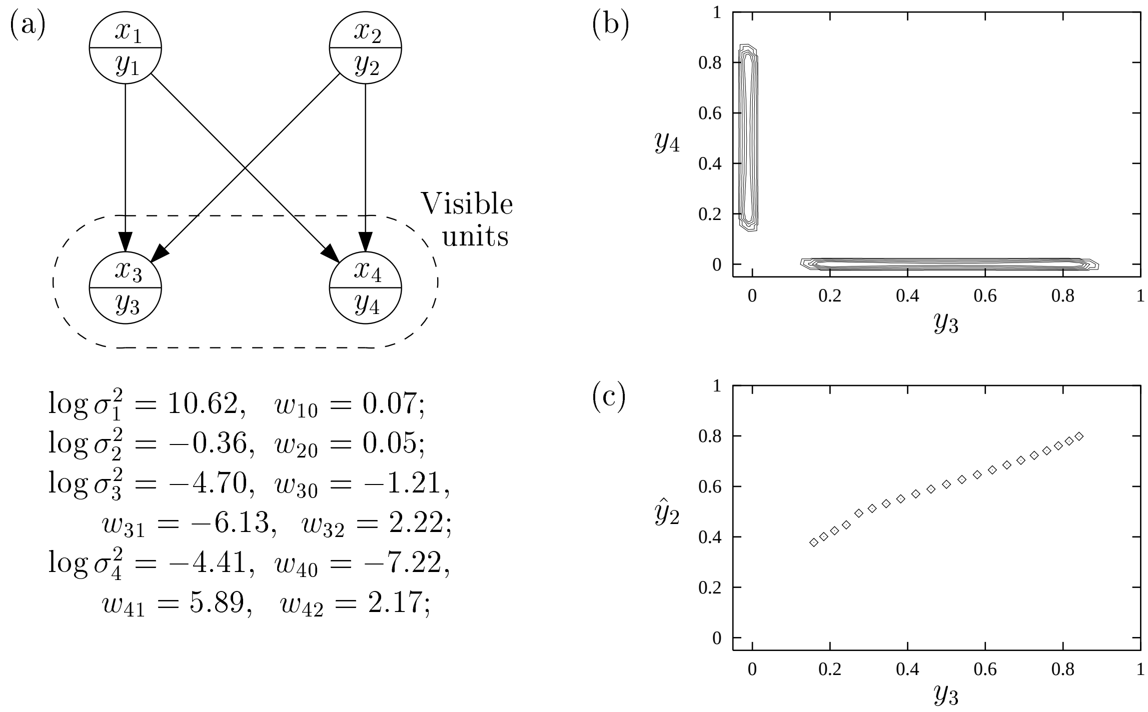
It can be shown quite easily that $M(\mu_j, s_j)$ has a closed-form solution:

$$M(\mu_j, s_j) = \Phi\Big(\frac{\mu_j}{\sqrt{1+s_j^2}}\Big). \tag{15}$$

As far as I know, $V(\mu_j, s_j)$ does not have a closed-form solution. However, since the coefficient in front of each $V(\mu_j, s_j)$ in equation 14 is positive, $F(Q, P)$ can be bounded from above by bounding each $V(\mu_j, s_j)$ from above, using

$$V(\mu_j, s_j) \le \Phi\Big(\frac{\mu_j}{\sqrt{1+s_j^2}}\Big)\Big[1 - \Phi\Big(\frac{\mu_j}{\sqrt{1+s_j^2}}\Big)\Big]\frac{s_j^2}{s_j^2 + \pi/2}. \tag{16}$$

Figure 3 shows the bound on the variance and the actual variance (computed by Monte Carlo) as functions of μ_j for several values of $\log s_j^2$. More important than tightness, the bound preserves the shape of the actual variance.



$$\log \sigma_1^2 = 10.62, \quad w_{10} = 0.07;$$
$$\log \sigma_2^2 = -0.36, \quad w_{20} = 0.05;$$
$$\log \sigma_3^2 = -4.70, \quad w_{30} = -1.21, \quad w_{31} = -6.13, \quad w_{32} = 2.22;$$
$$\log \sigma_4^2 = -4.41, \quad w_{40} = -7.22, \quad w_{41} = 5.89, \quad w_{42} = 2.17;$$

Figure 4: The network in (a) was obtained by applying Makov chain Monte Carlo parameter estimation to a data set whose density contours are shown in (b). Variational inference was applied to a sequence of data points taken from the lower "ridge". (c) shows the inferred value of y_2 for different positions y_3 along the ridge.

The upper bound on $F(Q,P)$ given by the variance bound is obviously not quadratic in the μ_j's or s_j's. I have implemented a variational optimization algorithm that uses conjugate gradients (Fletcher 1987) to minimize the upper bound on $F(Q,P)$.

Results

The variational inference algorithm described above was applied to a toy network that was estimated using the "slice sampling" Markov chain Monte Carlo method (Frey 1997). The trained network is shown in figure 4a and the density contours of the training data are shown in figure 4b.

Since σ_1^2 is very large, y_1 acts as a binary variable; in contrast, the mean and variance for unit 2 give y_2 a roughly uniform distribution on $(0,1)$. Given a visible vector (y_3, y_4), y_1 identifies to which "ridge" the data point belongs and y_2 reveals how far along the ridge the point lies. For various points along the lower ridge, figure 4c shows the inferred value $\hat{y}_2$ of unit 2 as a function of y_3. The inferred value was computed from $\hat{y}_2 = M(\mu_2, s_2) = \Phi(\mu_2/\sqrt{1+s_2^2})$. There is a monotonic and almost linear relationship between $\hat{y}_2$ and y_3.

Parameter estimation

Another justification for choosing $D(Q \parallel P)$ as the distance measure for variational inference follows from the fact that the inference statistics can be used to approximate maximum-likelihood parameter estimation. Recall that minimizing $D(Q \parallel P)$ with respect to the variational distribution is equivalent to minimizing $F(Q, P)$ in equation 11. Since the relative entropy $D(Q \parallel P)$ is always positive, $F(Q, P)$ is an upper bound on the negative log-probability of the data. So, maximum-likelihood parameter estimation can be approximated by alternately minimizing $F(Q, P)$ with respect to the variational distribution and the model parameters, in a fashion similar to the expectation maximization (EM) algorithm. This "variational EM" algorithm maximizes a lower bound $-F$ on the log-probability of the data.

As first pointed out by Jaakkola *et. al.* (1996) for their *binary* sigmoidal Bayesian networks, an advantage of using the cumulative Gaussian squashing function is that $F(Q, P)$ is quadratic in the weights, w_{ij}. The present *continuous* Bayesian network formulation retains this quadratic property in the w_{ij}'s (see equation 14). Furthermore, the values of the σ_i's do not influence the optimal w_{ij}'s, so that adjusting the model parameters is a matter of solving a linear system to obtain the w_{ij}'s, and then setting the σ_i's to their corresponding optimal values.

For the problem described in the previous section, I used data sets with 10,000 cases to train 40 models using the variational EM procedure described above. The initial weights for each model were drawn from a uniform distribution over [-0.1,0.1). The initial log-variances in each of 4 sets of 10 models were set to either -2.0, 0.0, 4.0 or 10.0. Singular-value decomposition (SVD) was used to solve the linear system described above. Not even a single model was correctly estimated; the estimated models produced probability densities that were roughly Gaussian without any higher-order structure.

Discussion

It is interesting that a variational inference method can work extremely well on a correctly estimated model (as shown above), and yet not be adequate for model estimation. One possible explanation for this behavior is that the SVD algorithm consistently finds poor log-probability local maxima in the model parameter space. However, in the above experiments the singular values were consistently of similar scale, indicating that the SVD algorithm was finding *the* maximum. Furthermore, in (Frey 1997), the models were properly estimated using *steepest descent*. A second explanation is that the conjugate gradient optimization finds local maxima in the variational distribution parameter space. A third explanation emerges from considering the implications of the restricted form of the variational distribution. Figure 5 shows a simplistic example of the $-F$ surface for a model with one observation. One horizontal axis represents configurations of the model parameters — each point on this axis identifies a unique model and consequently a unique distribution $P(\{x_j\}_{j \in H} | \{x_i\}_{i \in V})$. The other horizontal axis represents all possible distributions over the hidden variables — each point on this axis identifies a unique distribution $\Pi(\{x_j\}_{j \in H})$ over $\{x_j\}_{j \in H}$. The vertical axis gives $-F(\Pi, P) = -[D(\Pi \parallel P) - \log P(\{x_i\}_{i \in V})]$.

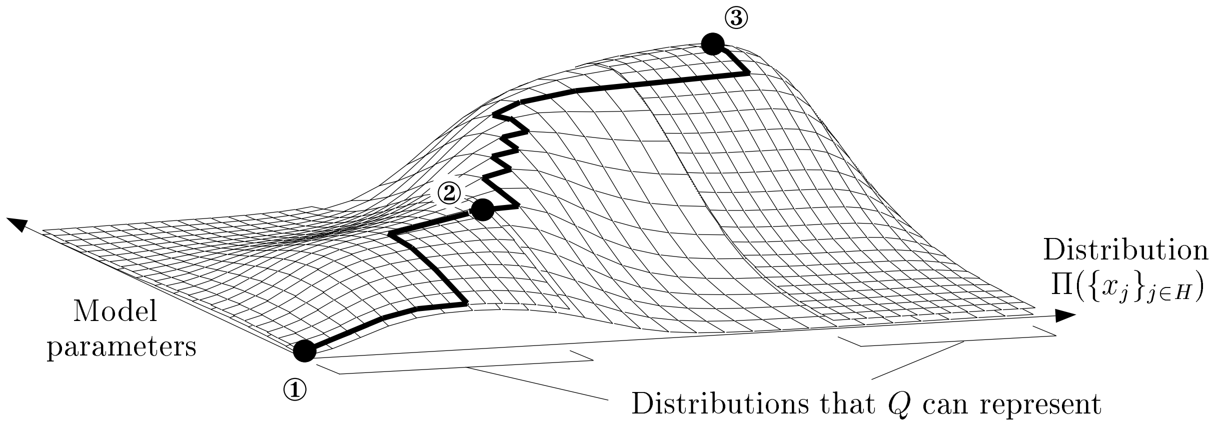


Figure 5: An example of $-F$ as a function of arbitrary distributions $\Pi(\{x_j\}_{j\in H})$ over the hidden variables (a generalized form of $Q(\{x_j\}_{j\in H})$) and the model parameters (which determine $P(\{x_j\}_{j\in H}|\{x_i\}_{i\in V})$). Starting from ①, the exact expectation maximimization (EM) algorithm follows the path shown and finds the maximum-likelihood solution at ③. Even though the variational parametric distribution Q is appropriate for the final solution at ③, it is does not represent intermediate distributions that are needed in the above path. As a result, the path is "blocked" at ② when variational EM is applied and the optimal solution at ③ is not found.

Consider the behavior of exact EM if the model parameters and the initial distribution Π are initialized at ①. Exact EM alternately maximizes $-F(\Pi, P)$ with respect to Π and the model parameters, as shown by the heavy line, until the optimum solution at ③ is found. Variational EM works in a similar fashion, except that only a subset of distributions can be represented by the parametric distribution Q, as shown in the figure. Notice that for both the initial model and the optimal model, the distribution Π that gives the maximal $-F(\Pi, P)$ can be represented by Q. However, when variational EM is applied starting at ①, the path taken by exact EM is "blocked" at ②. In order to find the best model, a sequence of intermediate distributions are needed that Q cannot represent.

There are two ways to broaden the spectrum of distributions that the variational distribution Q can represent. We can either abandon the product-form constraint or increase the complexity of each term in the product. The former approach will corrupt the locality of the interactions in $F(Q, P)$; *i.e.*, $F(Q, P)$ will not be decomposable into a sum of terms as in equation 12. The latter approach implies that $Q(x_i)$ becomes more complex, but $F(Q, P)$ can still be written as in equation 12. I am currently exploring distributions $Q(x_i)$ that are mixtures of Gaussians.

In general, it seems that through the choice of a variational distribution, the technique of variational inference offers plenty of elbow room for coming up with an inference algorithm that is both tractable and achieves a desired standard of quality.

Acknowledgments

I thank Zoubin Ghahramani for his comments on an earlier draft of this paper, and ITRC, IRIS and NSERC for financial support.

References

C. M. Bishop, M. Svensen and C. K. I. Williams 1996. EM optimization of latent-variable density models. In D. Touretzky, M. Mozer, and M. Hasselmo (editors), *Advances in Neural Information Processing Systems 8*, MIT Press, Cambridge, MA.

G. F. Cooper 1990. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence* **42**, 393-405.

P. Dayan, G. E. Hinton, R. M. Neal and R. S. Zemel 1995. The Helmholtz machine. *Neural Computation* **7**, 889-904.

E. Driver and D. Morrell 1995. Implementation of continuous Bayesian networks using sums of weighted Gaussians. In P. Besnard and S. Hanks (editors), *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA.

B. S. Everitt 1984. *An Introduction to Latent Variable Models.* Chapman and Hall, London, England.

B. J. Frey 1997. Continuous sigmoidal Bayesian networks trained using slice sampling. To appear in *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge, MA.

R. Fletcher 1987. *Practical methods of optimization.* John Wiley & Sons, New York, NY.

R. G. Gallager 1963. *Low-density parity-check codes.* MIT Press, Cambridge, MA.

Z. Ghahramani and M. I. Jordan, M.I. 1996. Factorial hidden Markov models. MIT Computational Cognitive Science Technical Report 9502.

D. Heckerman and D. Geiger, D. 1994. Learning Bayesian net- works: a unification for discrete and Gaussian domains. In P. Besnard and S. Hanks (editors), *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA.

G. E. Hinton, P. Dayan, B. J. Frey and R. M. Neal 1995. The wake-sleep algorithm for unsupervised neural networks. *Science* **268**, 1158-1161.

G. E. Hinton and T. J. Sejnowski 1986. Learning and relearning in Boltzmann machines. In D. E. Rumelhart and J. L. McClelland (editors), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations.* MIT Press, Cambridge, MA.

R. Hofmann and V. Tresp 1996. Discovering structure in continuous variables using Bayesian networks. In D. Touretzky, M. Mozer, and M. Hasselmo (editors), *Advances in Neural*

Information Processing Systems 8, MIT Press, Cambridge, MA.

T. Jaakkola, L. K. Saul and M. I. Jordan 1996. Fast learning by bounding likelihoods in sigmoid type belief networks. In D. Touretzky, M. Mozer and M. Hasselmo (editors), *Advances in Neural Information Processing Systems 8*, MIT Press, Cambridge, MA.

S. L. Lauritzen, A. P. Dawid, B. N. Larsen and H. G. Leimer 1990. Independence properties of directed Markov Fields. *Networks* **20**, 491-505.

S. L. Lauritzen and D. J. Spiegelhalter 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B* **50**, 157-224.

S. L. Lauritzen and N. Wermuth 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics* **17**, 31-57.

D. J. C. Mackay 1995. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research A* **354**, 73-80.

P. McCullagh and J. A. Nelder 1983. *Generalized Linear Models.* Chapman and Hall, London, England.

R. M. Neal 1992. Connectionist learning of belief networks. *Artificial Intelligence* **56**, 71-113.

R. M. Neal and G. E. Hinton 1993. A new view of the EM algorithm that justifies incremental and other variants. Unpublished manuscript available over the internet by ftp at `ftp://ftp.cs.utoronto.ca/pub/radford/em.ps.Z`.

R. M. Neal 1996. Markov chain Monte Carlo methods based on "slicing" the density function. In preparation.

J. Pearl 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA.

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

R. D. Shachter and C. R. Kenley 1989. Gaussian influence diagrams. *Management Science* **35**, 527-550.

D. J. Spiegelhalter and S. L. Lauritzen 1990. Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20**, 491-505.

R. Tibshirani 1992. Principal curves revisited. *Statistics and Computing* **2**, 183-190.