A variational approach to Bayesian logistic regression models and their extensions

Tommi S. Jaakkola and Michael I. Jordan
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139
{tommi,jordan}@psyche.mit.edu

October 31, 1996

Abstract

We consider a logistic regression model with a Gaussian prior distribution over the parameters. We show that accurate variational techniques can be used to obtain a closed form posterior distribution over the parameters given the data thereby yielding a posterior predictive model. The results are readily extended to (binary) belief networks. For belief networks we also derive closed form posteriors in the presence of missing values. 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

The Bayesian formalism is well suited for representing uncertainties in the values of variables, model parameters, or in the model structure. The formalism further allows ready incorporation of prior knowledge and the combination of such knowledge with statistical data (Bernardo & Smith 1994, Heckerman et al. 1995). The rigorous semantics, however, often comes with a sizable computational cost of evaluating multi-dimensional integrals. This cost precludes the use of exact 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 show how variational approximation techniques can restore the computational feasibility of the Bayesian formalism.

Variational techniques lead to deterministic approximations (or, in some cases, to exact results) and are used extensively in the physics literature (e.g., Sakurai 1985). These techniques transform the problem into an equivalent minimization (or maximization) problem by means of introducing extra variables known as variational parameters. The optimization of such parameters in turn often yields fixed point equations that can be solved iteratively. For the use of variational techniques in the context of graphical models see Saul et al. (1996) and Jaakkola and Jordan (1996).

Variational methods should be contrasted with sampling techniques (Neal 1994) that have become standard in the context of Bayesian calculations. While surely powerful in evaluating complicated integrals, sampling techniques do not guarantee monotonically improving approximations nor do they yield explicit bounds. It is precisely these issues that are the focus of the current paper.

The paper is organized as follows. First we develop a variational approximation method that allows the computation of posterior distributions for the parameters in Bayesian logistic regression

models. This is followed by a brief evaluation of the method's accuracy along with a comparison to other methods. We then extend the framework to belief networks, considering the case of incomplete data. Finally, we consider the dual of the regression problem – a density estimation 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

$$P(s|\mathrm{pa},\theta) = g\left((2s-1)\textstyle\sum_j \theta_j x_j\right) \tag{1}$$

where $g(x) = (1+e^{-x})^{-1}$ is the logistic function, s the binary response variable, and $\mathrm{pa} = \{x_1,\ldots,x_n\}$ 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

$$P(s|\mathrm{pa}) = \int P(s|\mathrm{pa},\theta)P(\theta)d\theta \tag{2}$$

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 $D^t = \{s^t, x_1^t,\ldots,x_n^t\}$ is a complete observation. To compute this posterior exactly, however, is not feasible.[1] It is nevertheless possible to find an accurate variational transformation of $P(s|\mathrm{pa},\theta)$ such that the desired posterior can be computed in closed form. Let us next introduce the transformation and show how the posterior can be computed based on a single observation D. We will see that under the variational approximation the parameter posterior remains Gaussian, and thus the full posterior can be obtained by sequentially absorbing the evidence from each of the observations.

The variational transformation we use is given by

$$\begin{aligned} P(s|\mathrm{pa},\theta) = g(X_s) &\geq g(\xi)\exp\left\{(X_s-\xi)/2 + \lambda(\xi)(X_s^2-\xi^2)\right\} &(3)\\ &= P(s|\mathrm{pa},\theta,\xi) &(4)\end{aligned}$$

where $X_s = (2s-1)\sum_j \theta_j x_j$ and $\lambda(\xi) = [1/2 - g(\xi)]/2\xi$. A direct maximization of the variational expression with respect to the variational parameter ξ recovers the original conditional distribution. The value of ξ at the maximum is simply X_s.

The posterior $P(\theta|D)$ can be computed by normalizing the left hand side of

$$P(s|\mathrm{pa},\theta)P(\theta) \geq P(s|\mathrm{pa},\theta,\xi)P(\theta) \tag{5}$$

given that this normalization is not feasible in practice we normalize the variational distribution instead. The variational distribution has the convenient property that it depends on the parameters θ only quadratically in the exponent (eq. 4). Consequently, as the prior distribution is a Gaussian with mean μ and covariance matrix Σ, computing the variational posterior – absorbing evidence – amounts to updating the mean and the covariance matrix. Omitting the algebra this update yields

$$\begin{aligned} \Sigma_{post}^{-1} &= \Sigma^{-1} + 2|\lambda(\xi)|xx^T &(6)\\ \mu_{post} &= \Sigma_{post}\left[\Sigma^{-1}\mu + (s-1/2)x\right] &(7)\end{aligned}$$

[1] Even without the prior distribution iterative schemes are needed and such methods are intractable in our case.

where $x = [x_1 \ldots x_n]^T$. Now, the posterior covariance matrix depends on the variational parameter ξ through $\lambda(\xi)$ and thus its value needs to be specified. We obtain ξ by optimizing the approximation in eq. (5). Using the fact that the approximation is in fact a lower bound we may devise a fast EM algorithm to perform this optimization (see appendix A). This leads to a closed form update for ξ given by

$$\xi^2 = E\left\{(\sum_j \theta_j x_j)^2\right\} = x^T \Sigma_{post}\, x + (x^T \mu_{post})^2 \tag{8}$$

where the expectation is taken with respect to $P(\theta|D, \xi^{old})$, the variational posterior distribution based on the previous value of ξ. Alternating between the ξ update and those of the parameters monotonically improves the posterior approximation of eq. (5). The convergence of this procedure is very fast; roughly only two iterations are needed. The accuracy of the resulting variational approximation is considered in the next section.

In summary the variational approach allows us to obtain the posterior predictive distribution

$$P(s|\mathrm{pa}, \mathcal{D}) = \int P(s|\mathrm{pa}, \theta) P(\theta|\mathcal{D}) d\theta \tag{9}$$

where the posterior distribution $P(\theta|\mathcal{D})$ comes from sequentially absorbing each (complete) observation D^t in the data set $\mathcal{D} = \{D^1, \ldots, D^T\}$. The predictive likelihoods $P(s^t|\mathrm{pa}_t, \mathcal{D})$ for any complete observation D^t have the form

$$\log P(s^t|\mathrm{pa}_t, \mathcal{D}) = \log g(\xi_t) - \xi_t/2 - \lambda(\xi_t)\xi_t^2 - \frac{1}{2}\mu^T \Sigma^{-1} \mu + \frac{1}{2}\mu_t^T \Sigma_t^{-1} \mu_t + \frac{1}{2} \log \frac{|\Sigma_t|}{|\Sigma|} \tag{10}$$

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 absorbing the evidence in D^t.

3 Accuracy of the variational method

Figure 1a compares the variational form of eq. (4) to the logistic function for a fixed value of ξ (here $\xi = 2$). We note that this is the optimized variational approximation in cases where $E\left\{(\sum_j \theta_j x_j)^2 | \xi = 2\right\} = 2^2$ since this condition is the fixed point of the update equation (8).

To get an indication of the quality of the variational approximation in the context of Bayesian calculations we numerically computed the approximation errors in the simple case where there is only one explanatory variable and the observation is $D = \{s = 1, x = 1\}$. Figure 1b shows the accuracy of the variational predictive likelihood as a function of different prior distributions. The evaluation of the posterior accuracy is deferred to the next section where comparisons are made to other related methods. In practice, we expect the accuracy of the posterior to be more important than that of the predictive likelihood since errors in the posterior run the risk of accumulating in the course of the sequential estimation procedure.

4 Comparison to other methods

Other sequential approximation methods have been proposed to yield closed form posterior parameter distributions in logistic regression models. The most closely related appears to be that of Spiegelhalter and Lauritzen (1990) (referred to as the S-L approximation in this paper). Their

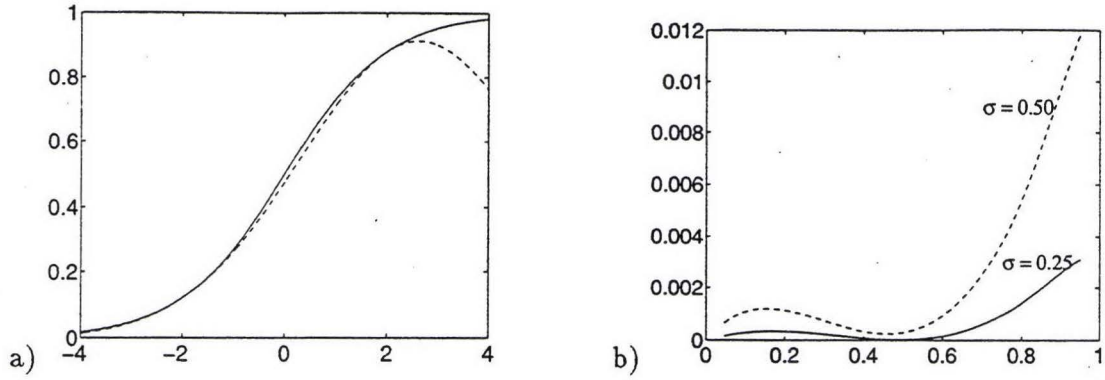

Figure 1: a) The logistic function (solid line) and its variational form (dashed line) when ξ is kept fixed at $\xi = 2$. b) The difference between the predictive likelihood $P(s = 1|\text{pa}) = \int g(\theta)P(\theta)d\theta$ and its variational approximation as a function of $g(\mu)$; here $P(\theta)$ is Gaussian with mean μ and variance σ^2.

method is based on making a local quadratic approximation to the complete log-likelihood centered at the prior mean μ (also known as the Laplace approximation). Similarly to the variational updates of eq. (6-7), the S-L approximation changes the prior distribution according to

$$\Sigma_{post}^{-1} = \Sigma^{-1} + \hat{p}(1-\hat{p})\, xx^T \tag{11}$$

$$\mu_{post} = \mu + (s-\hat{p})\Sigma_{post}\, x \tag{12}$$

where $\hat{p} = g(\mu^T x)$. Since there are no additional adjustable parameters in this approximation, it is simpler than the variational method. For the same reason, however, it can be expected to yield less accurate posterior estimates.

We compared the accuracy of the posterior estimates in the simple case where there is only one explanatory variable $x = 1$. The posterior of interest was $P(\theta|s = 1)$, computed for various settings of the prior mean μ and standard deviation σ. The correct second order statistics for the posterior were obtained numerically. Figures 2 and 3 illustrate the accuracy of the posterior for the two approximation methods. We used simple (signed) errors 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. Based on figures 2a and 3a the variational method yields more accurate estimates of the posterior means. When the prior variance is small (figure 2b), the S-L estimate of the posterior variance appears to be at least as good as the variational estimate. For larger prior variances, however, the S-L approximation degrades more rapidly. We note that the variational method consistently underestimates the true posterior variance – a fact that could also have been predicted theoretically (and could be used to refine the approximation). Finally, in terms of the KL-divergences between the approximate and true posteriors, the variational method seems to (slightly) outperform the S-L approximation, again the more clearly the larger the prior variance. This is shown in Figure 4.

5 Extension to belief networks

A belief network can be constructed from logistic regression models that define conditional probabilities of a variable given its parents[2]. The predictive joint distribution for this belief network

[2]The sets of parents for the variables must be consistent with some global ordering of the variables.

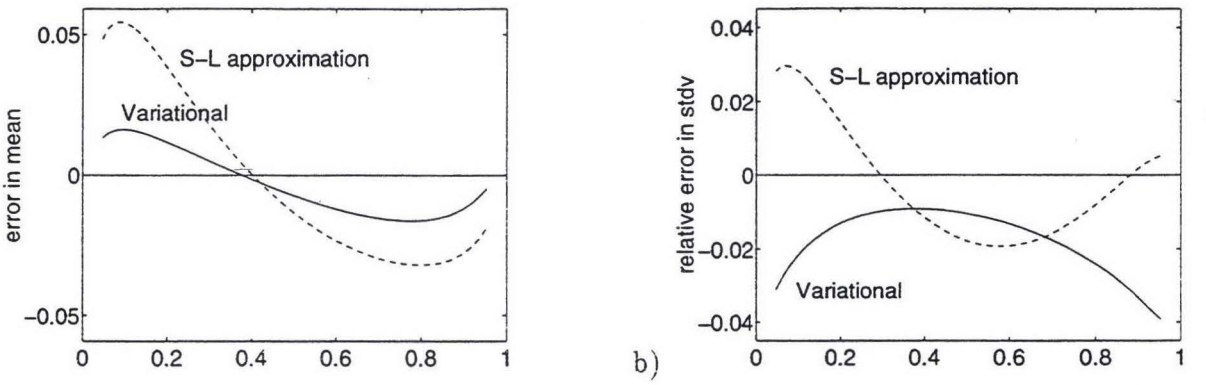

Figure 2: a) The errors in the posterior means as a function of $g(\mu)$, where μ is the prior mean. Here $\sigma = 1$ for the prior. b) The relative errors in the posterior standard deviations as a function of $g(\mu)$. $\sigma = 1$ for the prior distribution.

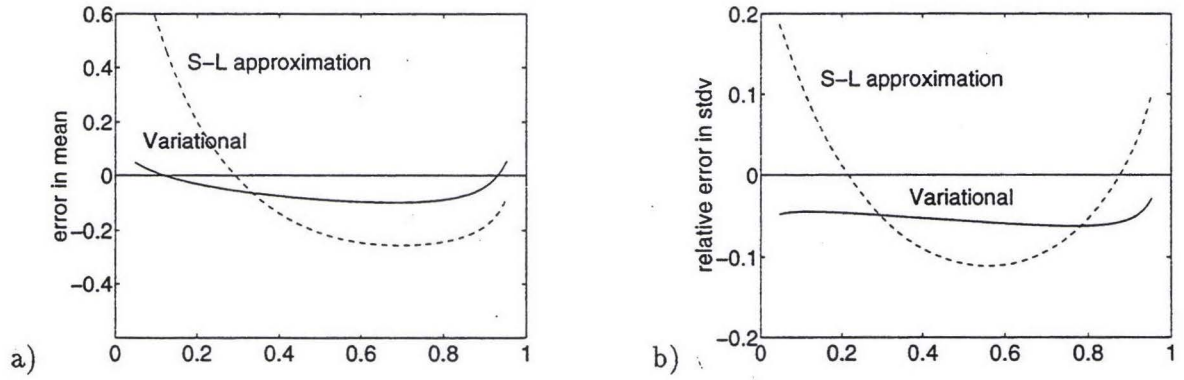

Figure 3: As figure 3 but now $\sigma = 2$ for the prior distribution.

takes the usual product form

$$P(s_1, \ldots, s_n) = \prod_i P(s_i|\mathrm{pa}_i) \tag{13}$$

We note that this is an extension of sigmoid belief networks (Neal 1992) due to the prior distributions over the parameters. In order to be able to use the techniques we have described for computing the posterior distributions on parameters in this setting, we assume that the observations are complete, i.e., contain a value assignment for all the variables in the network. Consequently, the parameter posteriors follow the factorization of the joint distribution and they can be computed separately for each conditional model – as before.

5.1 Incomplete cases: mean field inference

In many practical situations the assumption of complete cases is quite unrealistic. In the presence of missing values, however, the computation of posterior parameter distributions becomes quickly rather unwieldy because these posteriors now become dependent. This dependence arises from the need to sum over the possible configurations of the missing values weighted by their posterior probabilities. Introducing the variational transformation as in eq. (4) does not remove the ensuing dependence, nor does it allow the parameter posteriors to remain multivariate Gaussians (but a large mixture of them). Depending on the number of missing values, we may not even be able to

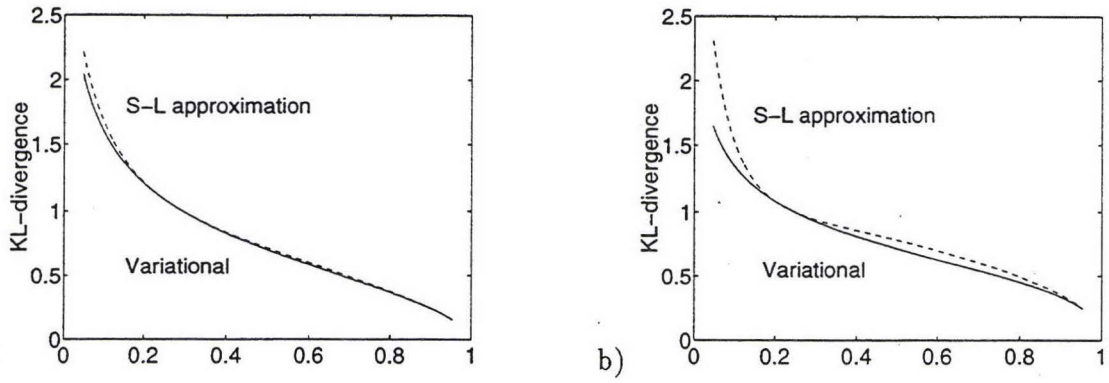

Figure 4: KL-divergences between the approximate and the true posterior distribution as a function of $g(\mu)$. a) $\sigma = 2$ for the prior. b) $\sigma = 3$. The two approximation methods have (visually) identical curves for $\sigma = 1$.

compute the sum over all the configurations; this can be exponentially costly, especially for densely connected networks.

We propose to deal with these problems by using mean field inference to fill in the missing values. The use of mean field constitutes an additional approximation, beyond the use of the variational transformation considered earlier. It nevertheless has the benefit of leaving the procedure for updating the parameter distributions relatively intact. Indeed, the approximate posterior distributions for the parameters remain Gaussians in this formulation, factorizing as with complete cases (see appendix B for details). The cost of such simplicity is the lack of correct dependencies across the parameter posteriors corresponding to different conditional models. While these posteriors remain probabilistically independent, their updates based on each new case are dependent (through mean field parameters). The detailed use of mean field with Bayesian parameter distributions is presented in appendix B. The sequential updating equations for the parameter distributions in this formulation are similar to eq. (6-7):

$$\Sigma_{post_i}^{-1} = \Sigma_i^{-1} + 2|\lambda(\xi_i)|\, E\left\{s_{\mathrm{pa}_i} s_{\mathrm{pa}_i}^T\right\} \tag{14}$$

$$\mu_{post_i} = \Sigma_{post_i}\left[\, \Sigma^{-1}\mu_i + E\left\{(s_i - 1/2)s_{\mathrm{pa}_i}\right\}\,\right] \tag{15}$$

where s_{pa_i} is the vector of parents of s_i, and the expectations are with respect to the mean field distribution. When the database cases are complete the expectations simply vanish. Unlike before the posterior distribution depends both on the variational parameter ξ and the mean field distribution. Again we can devise an EM algorithm to optimize these parameters iteratively (see appendix B.1).

6 The dual problem

The dual of the regression problem (eq. (1)) is found by switching the roles of the explanatory variables x and the parameters θ. In the dual problem, we have fixed parameters x and explanatory variables θ. Unlike before, distinct values of θ may explain different observations while the parameters x remain the same for all the observations, as shown in figure 5. In order to make the dual problem of figure 5b useful as a density model we generalize the binary output variables s to vectors $s = [s_1, \ldots, s_n]^T$ where each component s_i has a different set of parameters $x_i = [x_{i1} \ldots x_{im}]^T$ associated with it. The explanatory variables θ remain the same for all components. Consequently,

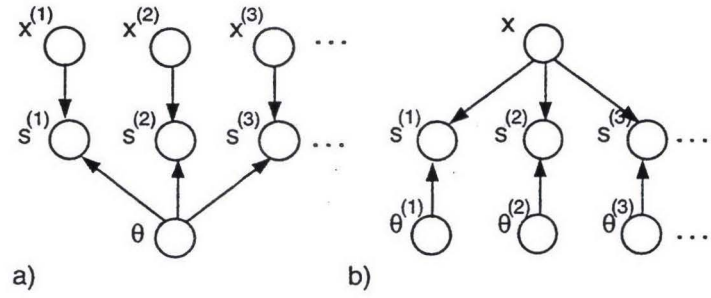

Figure 5: a) Bayesian regression problem. b) The dual problem.

the dual of the regression problem becomes a latent variable density model given by

$$P(s_1, \ldots, s_n|x) = \int \left[\prod_i P(s_i|\mathrm{pa}_i, \theta) \right] P(\theta) d\theta \tag{16}$$

where

$$P(s_i|\mathrm{pa}_i, \theta) = g\left((2s_i - 1) \textstyle\sum_j x_{ij} \theta_j \right) \tag{17}$$

In order to be able to use the EM algorithm for parameter estimation with this density model we again make use of the variational transformation of the conditional distributions. For each observation $D^t = \{s_1^t, \ldots, s_n^t\}$ this allows us to compute the posterior distribution $P(\theta|D^t, \xi^t)$ exactly. Analogously to the regression case the mean and the covariance of this posterior are given by

$$\Sigma_t^{-1} = \Sigma^{-1} + \sum_i 2|\lambda(\xi_i^t)| x_i x_i^T \tag{18}$$

$$\mu_t = \Sigma_t \left[\Sigma^{-1} \mu + \sum_i (s_i^t - 1/2) x_i \right] \tag{19}$$

The variational parameters ξ_i^t associated with each observation and output variable can be updated using eq. (8) assuming x is replaced with x_i. After obtaining these for all observations in the data set we may solve the M-step exactly and get

$$\Sigma \leftarrow \frac{1}{T} \sum_t \Sigma_t \tag{20}$$

$$\mu \leftarrow \frac{1}{T} \sum_t \mu_t \tag{21}$$

$$x_i \leftarrow A_i^{-1} b_i \tag{22}$$

where

$$A_i = \sum_t 2|\lambda(\xi_i^t)| (\Sigma_t + \mu_t \mu_t^T) \tag{23}$$

$$b_i = \sum_t (s_i^t - 1/2) \mu_t \tag{24}$$

We note finally that due to the variational transformation these updates result in a monotonically increasing *lower bound* on the log-likelihood of the observations.

7 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 likelihood of the observations is not guaranteed to increase after an iteration. We show here how to derived a monotonic, fast estimation procedure for logistic regression by making use of the variational transformation in eq. (4). Let us denote $X_t = (2s_t - 1)\sum_j \theta_j x_j^t$ and write the log-likelihood of the observations as

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_t \log P(s_t|\mathrm{pa}_t, \theta) = \sum_t \log g(X_t) \\ &\geq \sum_t \log g(\xi_t) + (X_t - \xi_t)/2 + \lambda(\xi_t)\left(X_t^2 - \xi_t^2\right) \\ &= \mathcal{L}(\theta, \xi) \end{aligned} \tag{25}$$

The variational lower bound is exact whenever $\xi_t = X_t$ 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-likelihood is a quadratic function of θ. The parameters θ that maximize $\mathcal{L}(\theta, \xi)$ are given by $\theta' = A^{-1}b$ where

$$A = \sum_t 2|\lambda(\xi_t)| x_t x_t^T \text{ and } b = \sum_t (s_t - 1/2)x_t \tag{26}$$

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

$$\mathcal{L}(\theta) = \mathcal{L}(\theta, \xi) \leq \mathcal{L}(\theta', \xi) \leq \mathcal{L}(\theta', \xi') = \mathcal{L}(\theta') \tag{27}$$

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

8 Conclusions

We have exemplified the use of variational techniques in a Bayesian inference problem. We found that variational methods can be employed to obtain closed form expressions that approximate the posterior distributions for the parameters in logistic regression and associated belief networks. Furthermore, our variational techniques lead to an exactly solvable EM algorithm for a type of latent variable density model—the dual of the regression problem.

Acknowledgments

This project was supported in part by NSF grant CDA-9404932, by a grant from the McDonnell-Pew Foundation, by a grant from Daimler-Benz Research, and by grant N00014-94-1-0777 from the Office of Naval Research. Michael I. Jordan is a NSF Presidential Young Investigator.

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** No. 3: 197.

T. Jaakkola and M. Jordan (1996). Computing upper and lower bounds on likelihoods in intractable networks. *Proceedings of the twelfth Conference on Uncertainty in Artificial Intelligence.*

T. Jaakkola and M. Jordan (1996). Recursive algorithms for approximating probabilities in graphical models. To appear in *Advances in Neural Information Processing Systems 9.*

P. McCullagh & J. A. Nelder (1983). *Generalized linear models.* London: Chapman and Hall.

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

R. Neal (1994). Bayesian Learning for Neural Networks. *PhD Thesis.* University of Toronto.

J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann: San Mateo.

J. Sakurai (1985). *Modern Quantum Mechanics.* Addison-Wesley.

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

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

A Optimization of the variational parameters

To optimize the variational approximation of eq. (5) in the context of an observation $D = \{s, x_1, \ldots, x_n\}$ 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

$$\int P(s|\mathrm{pa}, \theta) P(\theta) d\theta \geq \int P(s|\mathrm{pa}, \theta, \xi) P(\theta) d\theta \tag{28}$$

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

$$Q(\xi|\xi^{old}) = E\left\{\log P(s|\mathrm{pa}, \theta, \xi) P(\theta)\right\} \tag{29}$$

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

$$\frac{\partial}{\partial \xi} Q(\xi|\xi^{old}) = \frac{\partial \lambda(\xi)}{\partial \xi} \left[E\,(\sum_j \theta_j x_j)^2 - \xi^2 \right] = 0 \tag{30}$$

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

$$\xi^2 = E\,(\sum_j \theta_j x_j)^2 \tag{31}$$

By substituting ξ for ξ^{old} above, the procedure can be repeated. Each such iteration yields a better approximation in the sense of eq. (28).

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

B Parameter posteriors in the presence of missing values

Here we show how to obtain closed form expressions for the posterior parameter distributions in the presence of missing values. The simplicity comes at a cost of additional variational approximations that are needed to perform the fill in of the missing values. Let us start by considering the likelihood of the observed values D:

$$P(D) = \sum_{\{s\}\in\text{missing}} \prod_i \int P(s_i|\text{pa}_i,\theta_i)P(\theta_i)d\theta_i \tag{32}$$

$$\geq \int \left[\sum_{\{s\}\in\text{missing}} \prod_i P(s_i|\text{pa}_i,\theta_i,\xi_i) \right] \prod_i P(\theta_i)d\theta_i \tag{33}$$

$$\equiv \int P(D|\theta,\xi) \prod_i P(\theta_i)d\theta_i \tag{34}$$

where have introduced the variational transformation in (4). To be able to compute the variational probability of observations $P(D|\theta,\xi)$ we need to sum over the possible configurations of the missing values. This can be done efficiently by resorting to an additional bound:

$$\sum_s \exp\{\, f(s)\,\} \geq \exp\{\, E\{f(s)\} + H\,\} \tag{35}$$

where the expectation $E\{\cdot\}$ and the entropy H are calculated with respect to a variational distribution $\tilde{P}(s)$ over s. While the inequality or lower bound holds for arbitrary distributions $\tilde{P}$, the tightness of the bound depends on how closely $\tilde{P}(s)$ approximates $\exp\{f(s)\}$ (with normalization). Typically we choose a parametric form for $\tilde{P}$ and adjust the associated parameters so as to make the bound as tight as possible. If the variational distribution is chosen to be completely factorized the inequality corresponds to a mean field approximation (see e.g. Saul et al. 1996, or Jaakkola & Jordan 1996).

Recall that the variational conditional probabilities have the form

$$P(s_i|\text{pa}_i,\theta_i,\xi_i) = g(\xi_i) \exp\left\{(X_{s_i} - \xi_i)/2 + \lambda(\xi_i)(X_{s_i}^2 - \xi_i^2)\right\} \tag{36}$$

where $X_{s_i} = (2s_i - 1)\sum_j \theta_{ij}s_j$. Using now the variational lower bound technique of eq. (35) to transform the summation over the missing values we get

$$P(D|\theta,\xi) = \sum_{\{s\}\in\text{missing}} \prod_i P(s_i|\text{pa}_i,\theta_i,\xi_i) \tag{37}$$

$$\geq \exp\{H\} \prod_i g(\xi_i) \exp\left\{(E\{X_{s_i}\} - \xi_i)/2 + \lambda(\xi_i)(E\{X_{s_i}^2\} - \xi_i^2)\right\} \tag{38}$$

$$\equiv P(D|\theta,\xi,\tilde{p}) \tag{39}$$

where $\tilde{p}$ denotes the parameters of the variational distribution $\tilde{P}$. In order to quantify the expectations above we need to specify the parametric form of $\tilde{P}$. For simplicity we choose $\tilde{P}$ from the family of completely factorized (mean field) distributions

$$\tilde{P}(s) = \prod_i \tilde{p}_i^{s_i}(1 - \tilde{p}_i)^{1-s_i} \tag{40}$$

The relevant expectations in eq. (38) now become:

$$E\{X_{s_i}\} = (2\tilde{p}_i - 1)\sum_j \theta_{ij}\tilde{p}_j \tag{41}$$

$$E\{X_{s_i}^2\} = (\sum_j \theta_{ij}\tilde{p}_j)^2 + \sum_j \theta_{ij}^2 \tilde{p}_j(1-\tilde{p}_j) \tag{42}$$

$$H = \sum_i H(\tilde{p}_i) \tag{43}$$

where $H(\cdot)$ is the binary entropy function. In the simplified notation above, $\tilde{p}_i = s_i$ whenever s_i is observed.

Let us now consider how to compute the posterior parameter distribution. Crucial to the simplicity of this computation is the assumption that the additional variational parameters $\tilde{p}_i$ are functionally independent of the parameters θ_{ij}. This assumption can indeed be made; it only affects the tightness of the additional bound. Consequently, $P(D|\theta,\xi,\tilde{p})$ above becomes a quadratic function in the exponent with respect to θ (see the explicit forms of the expectations above). This property guarantees that the posterior parameter distribution remains in the Gaussian family (as in the logistic regression case) if computed by normalizing the right hand side of

$$P(D|\theta)P(\theta) \geq P(D|\theta,\xi,\tilde{p})P(\theta) \tag{44}$$

The mean and the covariance of this posterior can be shown to be

$$\Sigma_{post_i}^{-1} = \Sigma_i^{-1} + 2|\lambda(\xi_i)|\, E\left\{s_{\mathrm{pa}_i} s_{\mathrm{pa}_i}^T\right\} \tag{45}$$

$$\mu_{post_i} = \Sigma_{post_i}\left[\,\Sigma^{-1}\mu_i + E\left\{(s_i - 1/2)s_{\mathrm{pa}_i}\right\}\,\right] \tag{46}$$

where s_{pa_i} is the vector of parents of s_i and the expectations are taken with respect to the $\tilde{P}$ distribution. The factorization assumption makes these expectations easy to compute.

B.1 Optimization of the variational parameters

The posterior distribution depends on the variational parameters ξ and $\tilde{p}$. To optimize these parameters we may proceed as in appendix A and devise an EM algorithm to maximize the right hand side of

$$\log P(D) \geq \log \int P(D|\theta,\xi,\tilde{p})P(\theta)d\theta \tag{47}$$

Since the optimization of ξ can be carried out as in appendix A we will not repeat the procedure here but instead concentrate on $\tilde{p}$. The complete log-likelihood is a quadratic function of θ, and thus we can compute

$$Q(\xi,\tilde{p}|\xi^{old},\tilde{p}^{old}) = E\left\{\log P(D|\theta,\xi,\tilde{p})P(\theta) \;\middle|\; \xi^{old},\tilde{p}^{old}\right\} \tag{48}$$

in closed form. The expectation here is over the posterior distribution $P(\theta|D,\xi^{old},\tilde{p}^{old})$. Taking the derivative with respect to each $\tilde{p}_\alpha$ gives

$$\frac{\partial}{\partial\tilde{p}_\alpha}Q(\xi,\tilde{p}|\xi^{old},\tilde{p}^{old}) = \mu_{post_\alpha}^T\tilde{p}_{pa_\alpha} + \sum_{i\in ch_\alpha} 2\lambda(\xi_i)\, M_{i\alpha}\,\tilde{p}_{pa_i}\Big|_{\tilde{p}_\alpha=1/2} - \log\frac{\tilde{p}_\alpha}{1-\tilde{p}_\alpha} = 0 \tag{49}$$

where $M_{i\alpha j} = (\Sigma_{post_i} + \mu_{post_i}\mu_{post_i}^T)_{\alpha j}$ mediates the influence of other parents of i on α. Here the subscript *post* refers to the parameters (mean and the covariance) of $P(\theta|\xi^{old},\tilde{p}^{old})$. We note that

the derivative depends on $\tilde{p}_\alpha$ only through the log term and therefore the maximizing $\tilde{p}_\alpha$ is readily found. We obtain

$$\tilde{p}_\alpha \leftarrow g\left(\mu_{post_\alpha}^T \tilde{p}_{pa_\alpha} + \sum_{i \in ch_\alpha} 2\lambda(\xi_i)\, M_{i\alpha}\, \tilde{p}_{pa_i} \Big|_{\tilde{p}_\alpha = 1/2}\right) \tag{50}$$

Iteratively solving for each of the variational "means" $\tilde{p}_\alpha$ in this manner leads to a monotonically increasing lower bound on the log-likelihood of the observations. The updates of $\tilde{p}$ may be interlaced with those of ξ, and the posterior distribution over θ need not be recomputed after each update of the variational parameters.