Variational Methods in Bayesian Deconvolution

K. Zarb Adami
Cavendish Laboratory, University of Cambridge, UK

This paper gives an introduction to the use of variational methods in Bayesian inference and shows how variational methods can be used to approximate the intractable posterior distributions which arise in this kind of inference. The flexibility of these approximations allows us to include positivity constraints when attempting to infer hidden pixel intensities in images. The approximating posterior distribution is then optimised by minimising the Kullback-Leibler divergence between it and the true distribution. Unlike traditional methods such as Maximum Likelihood or Maximum-A-Posteriori methods, the variational approximation is immune to overfitting, since the sensitivity of the approximation is towards probability mass rather than probability density. The results show that the present algorithm is successful in interpolation and deconvolution problems.

1. INTRODUCTION

1.1. Measurement

Before the analysis of data is considered, it is worth discussing the measurement process by which the data is collected. Measurement involves an interaction between the instrument and the environment, possibly involving uncertainty in the obtained result. This uncertainty takes on many different forms including both systematic and random effects. The most general form of writing down a measurement is as follows:

$$D = R(\Theta) + \nu \tag{1}$$

where D is the obtained measurement, R is the response function of the instrument as a function of the parameters Θ we set out to measure, and ν is the noise or uncertainty introduced by the environment. Data analysis involves the treatment of the collected measurements to extract the required parameters.

The response function is often complicated, so that even without noise Equation (1) cannot be directly inverted to obtain the parameters. This means that an approximate method is required to form a suitable pseudo-inverse. Since many pseudo-inverses are possible this inversion process is ill-determined and a systematic way of choosing the correct one is required. Throughout this paper we employ a probabilistic solution to this general inference problem. R.A. Fisher discusses three aspects of valid inference: (1) model specification, (2) estimation of model parameters, and (3) estimation of precision. Model specification can be further subdivided into two main categories: the formulation of a set of candidate models, and the selection of a model (or small number of models) to be used in performing inference.

1.2. Model Selection

Choosing the best model typically requires a balance between minimising a cost function, the most ubiquitous one being the chi-squared statistic χ^2, and a regularising function, a common one being the entropy $-\sum_i p_i \log p_i$. The correct balance is deduced through the use of Bayes' theorem:

$$P(\Theta|\mathcal{D},\mathcal{H}) = \frac{\mathcal{L}(\mathcal{D}|\Theta,\mathcal{H}) \times P(\Theta|\mathcal{H})}{P(\mathcal{D}|\mathcal{H})} \tag{2}$$

where Θ are the parameters we are hoping to infer, $\mathcal{D}$ are the data we have collected and $\mathcal{H}$ is the model under scrutiny. The likelihood function, $\mathcal{L}(\mathcal{D}|\Theta,\mathcal{H})$, contains all the parameters we are seeking to infer and is a functional description of the relationship between the parameters and the data. The prior, $P(\Theta|\mathcal{H})$, contains all the knowledge available to the experimenter before the measurement is performed. It reflects any assumptions made by the experimenter and it can contain constraints on the range in which the data should be. The posterior distribution, $P(\Theta|\mathcal{D},\mathcal{H})$, quantifies the belief in the parameters inferred from the data. The denominator of Equation (2) is not merely a normalisation constant, as a further application of Bayes' theorem shows:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H}) \times P(\mathcal{H})}{P(\mathcal{D})} \tag{3}$$

The evidence ($P(\mathcal{D}|\mathcal{H})$) is crucial for model selection since Equation (3) is merely the posterior probability of the model given the data. Unfortunately, evaluation of the evidence is a hard problem, involving high-dimensional integrals over parameter space [6]. Although sampling methods have been more popular recently, this paper focuses on a variational method for evaluating a bound on the evidence value and using this bound to perform inference.

2. VARIATIONAL INFERENCE

Variational inference finds itself between the Laplace approximation and sampling methods. At one extreme, Laplace's method approximates the integrand by a Taylor expansion of the logarithm of the

posterior around its peak. This method is unwieldy in high-dimensional problems since a large matrix of cross-derivatives is required. At the other extreme, we can approximate the evidence using numerical techniques such as Markov Chain Monte Carlo methods (MCMC) [4]. These are however very computationally intensive.

Variational inference attempts to approximate the integrand until the integral becomes tractable. The general idea, following [7], is to bound the integrand from above or below, reducing the integration problem to an optimisation problem, i.e. making the bound as tight as possible. No parameter estimation is required and the quality of the integral is optimised directly. The Kullback-Leibler cross-entropy is used as a measure of the disparity between the true and approximate posterior ($Q(\Theta)$), and quantifies the loss of information incurred through the approximation.

$$D_{KL}(Q||P) = \int_\Theta Q(\Theta) \log\left[\frac{Q(\Theta)}{P(\Theta|\mathcal{D},\mathcal{H})}\right] d\Theta \quad (4)$$

This can be re-arranged to:

$$\begin{aligned} C_{KL}(Q||P) &= D_{KL}(Q||P) - \log P(\mathcal{D}|\mathcal{H}) \\ &= \int_\Theta Q(\Theta) \log\left[\frac{Q(\Theta)}{P(\mathcal{D}|\Theta,\mathcal{H})P(\Theta|\mathcal{H})}\right] d\Theta \\ &\geq -\log P(\mathcal{D}|\mathcal{H}) \end{aligned} \quad (5)$$

so that the minimum of C_{KL} corresponds to the optimum approximating distribution which provides a lower bound for $\log P(\mathcal{D}|\mathcal{H})$.

This approach allows flexibility in specifying the prior and provides a deterministic way of obtaining a bound on the evidence value. A strength of this approximation lies in its sensitivity to probability mass rather than probability density.

3. INTERPOLATION

Consider the problem of interpolating a curve through a set of points, so that the generative model for our data is:

$$\mathcal{D}_i = \sum_{n=1}^{N} w_n f_{ni} + \nu_i \quad (6)$$

where $\mathcal{D}$ is the observed data, $\mathbf{w}$ is a vector of parameters and $\mathbf{f}$ is a matrix of basis functions. If we assume our noise model to be gaussian with inverse variance γ, we can immediately write down the likelihood:

$$\mathcal{L}(\mathcal{D} \mid w, \gamma, \mathcal{H}) = \prod_{i=1}^{I} \mathcal{G}\left(\mathcal{D}_i \mid \sum_{n=1}^{N} w_n f_{ni}, \gamma\right) \quad (7)$$

Bayes' theorem now demands we specify prior distributions for the parameters we are trying to infer, namely $\mathbf{w}$ and γ. The variational method provides the freedom to choose any analytical form for our priors, and we choose priors conjugate to the posterior distribution to provide a suitable analytical approximation. Furthermore, the resulting approximation will have the same form as the prior, so that a set of posterior distributions from one data set could be used as a prior distribution for a new set of data. The resulting prior distributions for $\mathbf{w}$ and γ are [5]:

$$\begin{aligned} p(\mathbf{w}|\mathcal{H}) &= \mathcal{G}(\mathbf{w}|0, a^{(w)}\mathbf{I}) \\ p(\gamma|\mathcal{H}) &= \text{Gamma}(\gamma|a^{(\gamma)}, b^{(\gamma)}) \end{aligned} \quad (8)$$

Both Maximum Likelihood and Variational methods are now used to perform inference on this problem.

3.1. Maximum Likelihood

By differentiating the likelihood function with respect to the parameters we are attempting to infer, we have:

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathcal{D} \mid w, \gamma, \mathcal{H})}{\partial \mathbf{w}} &= 0 \\ \Rightarrow \mathbf{w} = [\mathbf{f}\mathbf{f}^T]^{-1}\mathbf{f}\mathcal{D} \end{aligned} \quad (9)$$

and similarly we obtain

$$\gamma^{-1} = \frac{1}{J}\sum_{i=1}^{J}\left(\mathcal{D}_j - \sum_{n=1}^{N} w_n f_{nj}\right)^2 \quad (10)$$

where J is the number of data points in the time series.

3.2. Variational Learning

With the variational method, the optimal approximate posterior distributions are found to be [5]:

$$Q(\mathbf{w}) = \mathcal{G}(\mathbf{w}|\widehat{\mathbf{w}}, \widetilde{\mathbf{w}}) \quad (11)$$

$$Q(\gamma) = \text{Gamma}(\gamma|\overline{a}^{(\gamma)}, \overline{b}^{(\gamma)}) \quad (12)$$

where the parameters obey

$$\widetilde{\mathbf{w}} = a^{(w)}\mathbf{I} + \langle\gamma\rangle_Q \mathbf{f}\mathbf{f}^T \quad (13)$$

$$\widetilde{\mathbf{w}} = \widetilde{\mathbf{w}}^{-1}\langle\gamma\rangle_Q \mathbf{f}\mathcal{D} \quad (14)$$

$$\overline{a}^{(\gamma)} = a^{(\gamma)} + \frac{1}{2}\sum_{j=1}^{J}\left\langle\left(\mathcal{D}_j - \sum_{n=1}^{N} w_n f_{nj}\right)^2\right\rangle_Q \quad (15)$$

$$\overline{b}^{(\gamma)} = b^{(\gamma)} + \frac{J}{2} \quad (16)$$

where $\mathbf{I}$ is the identity matrix.

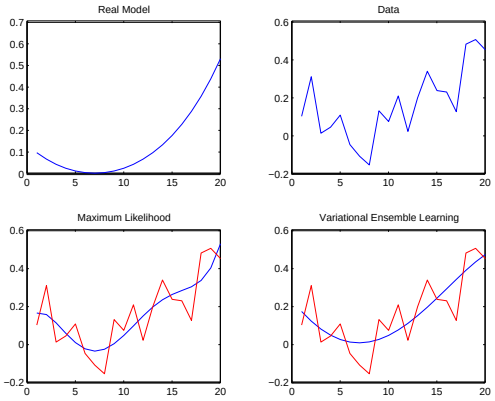

Figure 1: The Maximum Likelihood and Variational Learning algorithms for the interpolation model.

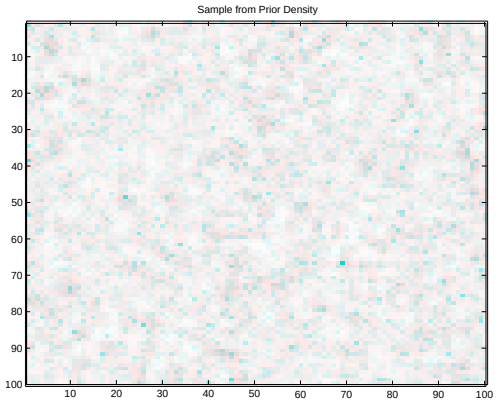

Figure 2: Samples from a mixture of Laplacians. This prior distribution favours sparse images.

3.3. Results

The results displayed in the figure were obtained by trying to approximate the curve $y(x) = \exp(x-0.3)-1$ after gaussian noise has been added to it. The true curve and the data are displayed in the top two panels, while the inferred curves are plotted in blue in the bottom two panels.

The two inference algorithms are similar in computational complexity, though it is clear from the bottom left panel that the Maximum Likelihood method is over-fitting the data, whereas the Variational method seems to approximate the true curve accurately. Also the variational method returns a distribution over possible interpolation curves, unlike the maximum likelihood method which just returns one.

4. DECONVOLUTION

Very often, Equation 1 can be written as a convolution process, where the response function is convolved with the true source distribution, such that Equation 1 becomes linear and of the form:

$$\mathcal{D} = \mathcal{A} * s + \nu \tag{17}$$

where $\mathcal{A}$ is the response or beam function of the apparatus, s is the true source-distribution (pixel intensity in an image case) and ν is the noise system. This means that we now need to specify prior distributions over the source distribution of the pixels in our image. In the next section we discuss the case in which the beam function is also unknown.

Pixel intensity is a positive quantity and our prior distribution should reflect this fact [3]. A suitable distribution is the Laplace distribution. To provide further flexibility, we can model the intensities as a mixture of Laplacian distributions given by:

$$p(s_{ij}) = \begin{cases} \sum_{\alpha=1}^{N_\alpha} \frac{\pi_\alpha}{b_\alpha} \exp\left(-\frac{s_{ij}}{b_\alpha}\right) & s_{ij} \geq 0 \\ 0 & s_{ij} < 0 \end{cases} \tag{18}$$

where ij represents the ij^{th} pixel in the image. Priors over the hyper-parameters π_α and b_α need to be specified in order to ensure sufficiently-broad priors over the pixel intensity. For b_α, a scale-invariant prior is selected so that:

$$p(\ln b_\alpha) = 1 \tag{19}$$

while, since π_α represents the fraction of mixture α present, we use a Dirichlet prior of the form:

$$p(\pi_\alpha) \propto \delta\left(\prod_{\alpha=1}^{N_\alpha} \pi_\alpha - 1\right) \prod_{\alpha=1}^{N_\alpha} \pi_\alpha \tag{20}$$

If we now assume the additive noise is gaussian, we can immediately write down the likelihood function:

$$\mathcal{L}(\mathcal{D}|s, \beta_\sigma) = \prod_{ij} \mathcal{G}(\mathcal{D}_{ij}|\hat{\mathcal{D}}_{ij}; \beta_\sigma^{-1}) \tag{21}$$

where β_σ is the inverse variance of the noise. Since we do not know this quantity, we must also assign it a prior. A scale invariant prior of the form:

$$p(\ln \beta_\sigma) = 1 \tag{22}$$

is used. We can now easily form the posterior distribution over the required parameters, namely s_{ij} and β_σ. Following the variational method, we now suppose that the posterior distributions are separable, so that we can write the approximate tractable distributions as:

$$Q(\Theta|\mathcal{H}) = Q(s, \beta_\sigma|\mathcal{H}) = \prod_{ij} (Q(s_{ij})) \times Q(\beta_\sigma|\mathcal{H}) \tag{23}$$

This assumption allows us to separate out the terms in the cost function so that it can be written as a sum of simple individual terms. Again, a specific form of these posterior distributions is not required, since the forms which optimise the cost function subject to the separable form and normalisation conditions can be found via the variational method. Following [6], each distribution can then be updated in turn using the current estimates for all the other distributions.

4.1. Example

As an example we consider the toy problem of deconvolving some text. Starting with the source distribution displayed in the bottom left panel, we convolve it with a Gaussian beam function and add some gaussian noise to it to obtain the panel in the bottom right of the figure.

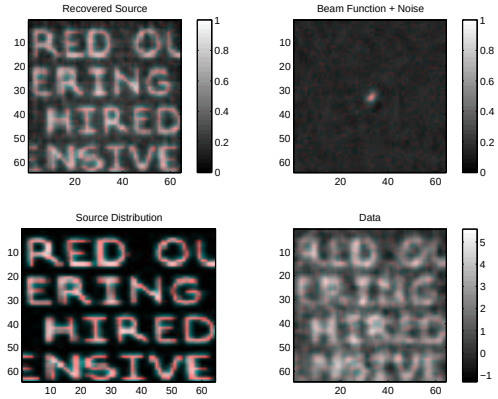

Figure 3: Deconvolution of a noisy image using the variational method with a mixture of Laplacians as a prior.

By learning the source distribution through the variational method, we obtain the result in the top left panel of the figure. However, in astrophysical problems it is sometimes the case that the beam function too is unknown. This problem is common in optical interferometry, where the incoming wavefronts are affected by atmospheric turbulence.

5. BLIND DECONVOLUTION

If the beam function is unknown too, we need to include it in the set of parameters we are trying to infer. In order to tackle the blind problem, we assume (following [2]) that (a) the beam function is smaller than the underlying source distribution and, more importantly, that (b) the beam function is **independent** of the source distribution. We now write down the blind deconvolution problem, following [6], such that:

$$\mathcal{D}_{ij} = \sum_{k=-K}^{K} \sum_{l=-K}^{K} A_{kl} s_{i-k,j-l} + \nu_{ij} \qquad (24)$$

where A_{kl} is an element of the beam matrix, which we have assumed to be square and of side $2K$. In evaluating the above sum we assume that the source distribution outside the defined extent of the image is zero. In addition to our deconvolution priors we must now specify a prior over the elements of the beam matrix. In order to respect positivity a Laplacian prior is selected so that:

$$p(A_{kl}) = \begin{cases} \beta_a \exp\left(-\beta_a A_{kl}\right) & A_{kl} \geq 0 \\ 0 & A_{kl} < 0 \end{cases} \qquad (25)$$

As in the previous section we use a scale invariant prior for β_a. If we again approximate the posterior distribution by separable distributions for each parameter, we can derive the update equations for the approximate posteriors, following [5], and then iteratively update the posterior distributions. Below is an example in which the beam matrix is unknown and is inferred using the above method.

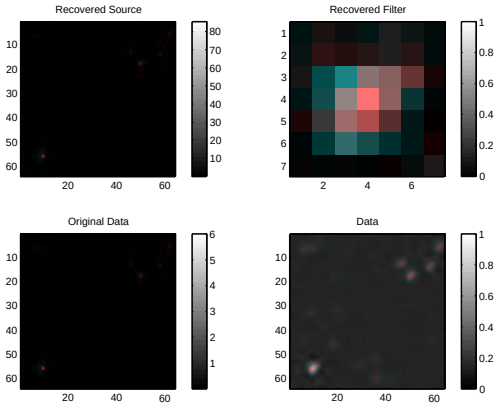

Figure 4: Blind deconvolution of a noisy image using the variational method.

As seen in the figure above, we can use the variational method to successfully infer both the beam function and the underlying source distribution. Throughout the previous two sections, we have assumed that pixel intensities are independent and have neglected any intrinsic correlations which may exist. As a further step, one might model an image as an independent set of pixels convolved with another unknown beam matrix, and priors over this beam matrix could be specified and inferred.

6. CONCLUSIONS

This paper has shown how variational methods can be used to perform valid inference by approximating

the true posterior distribution by tractable solutions. The strength of variational methods lies in the reduction of a high-dimensional integration problem to a high-dimensional optimisation problem. The results presented in this paper demonstrate that the variational method is useful in both deconvolution and blind deconvolution problems as well as other inference problems. Work in combining the variational method with MCMC methods is continuing with the aim of using the variational approximation to speed up MCMC navigation through posterior space.

Acknowledgments

The author thanks ST Microelectronics for financial support of this project as well as Steve Gull, David Mackay and Michael Hobson for useful discussions.

References

[1] H.Attias, "Blind Source Separation and Deconvolution: The dynamic component analysis algorithm.", Neural computation **10**, pp 1373-1424, 1998.

[2] H.Attias, "Independent factor analysis.", Neural computation **11**, pp 803-805, 1998.

[3] S.F. Gull and G.J. Daniell, "Image reconstruction from incomplete and noisy data.", Nature **272**, pp 686-690, 1978.

[4] J. Skilling, "Bayesys3 Users' Manual.", 2003.

[5] J. Miskin, "Ensemble Learning for Independent Component Analysis", Ph.D. Thesis, University of Cambridge, December 2000.

[6] J. Miskin, D.J.C. Mackay, "Ensemble Learning for Blind Image Separation and Deconvolution", Chapter 8 Independent Components Analysis: Principles and Practice, Cambridge University Press

[7] T.P. Minka,"Using lower bounds to approximate integrals", Technical Report, Medai Lab, Massachusetts Institute of Technology, June 2001