Stat260: Bayesian Modeling and Inference **Lecture Date: March 31, 2010**

Lecture 16

Lecturer: Michael I. Jordan *Scribe: Samitha Samaranayake*

1 Laplace approximation review

In the previous lecture we discussed the Laplace approximation as a general way to approach marginalization problems. The basic idea was to approximation an integral of the following form:

$$I(t) = \int e^{-Nh(x)} dx, \tag{1}$$

where N is typically the number of data points. After performing a Taylor series expansion of both $h(x)$ and the exponential function and evaluating some elementary integrals, we showed that the following approximation of $I(t)$ can be derived.

$$I(N) = e^{-Nh(\hat{x})}\sqrt{2\pi}\sigma N^{-1/2}\left(1 - \frac{h_4(\hat{x})\sigma^4}{8N} + \frac{5h_3^2(\hat{x})\sigma^6}{24N}\right) + O(1/N^2), \tag{2}$$

where $\sigma^2 = 1/h_2(\hat{x})$ and $\hat{x} = \text{argmin}_x\, h(x)$. If $\hat{x}$ can not be determined analytically, it is typically approximated with some value $\tilde{x}$ such that the approximation error of $\hat{x} - \tilde{x}$ is within a factor of $O(1/N)$. For example, in a Bayesian application $-h(x)$ can be the likelihood times a prior and $\hat{x}$ is then the maximum a posteriori probability (MAP). As N gets large, the MAP approaches the maximum likelihood (ML) estimate, so we can approximate $\hat{x}$ with the MLE and still obtain a rigorous accuracy bound.

2 Multivariate Laplace approximation

The multivariate case is derived in exactly the same way as the univariate case was derived in lecture 15. The only difference being that we perform a multivariate Taylor series expansion and get a multivariate Gaussian integral. Letting x denote a d-dimensional vector and $h(x)$ a scalar function of x, we obtain:

$$\int e^{-Nh(x)} dx \approx e^{-Nh(\hat{x})}(2\pi)^{d/2}|\Sigma|^{1/2}N^{-d/2}, \tag{3}$$

where $\Sigma = (D^2h(\hat{x}))^{-1}$ is the inverse of the Hessian of h evaluated at $\hat{x}$. This expansion is accurate to order $O(1/N)$, since we only consider the first order terms of the Laplace approximation. However, as in Eq. (2), the expansion can be continued to obtain an accuracy of order $O(1/N^2)$.

3 Marginal likelihood

One application of the Laplace approximation is to compute the marginal likelihood. Letting M be the marginal likelihood we have,

$$\begin{aligned} M &= \int P(X|\theta)\pi(\theta)\, d\theta \\ &= \int \exp\left\{-N\left(-\frac{1}{N}\log P(X|\theta) - \frac{1}{N}\log \pi(\theta)\right)\right\} d\theta \end{aligned} \tag{4}$$

where, $h(\theta) = -\frac{1}{N}\log P(X|\theta) - \frac{1}{N}\log \pi(\theta)$. Using the Laplace approximation up to the first order as in Eq. (3) we get,

$$M \approx P(X|\hat{\theta})\pi(\hat{\theta})(2\pi)^{d/2}|\Sigma|^{1/2}N^{-d/2} \tag{5}$$

This approximation is used for example in model selection, where computing the marginal likelihood analytically can be hard unless there is conjugacy. Computing the Laplace approximation requires finding the maximum a posteriori probability $\hat{\theta} = \text{argmax}_\theta\, -h(x)$, which can be done using a standard method such as gradient search. It also requires computing the second derivative matrix and inverting it to obtain Σ. This is usually the harder quantity to calculate.

4 Bayesian information criterion (BIC) score

The Bayesian information criterion[1] score tries to minimize the impact of the prior as much as possible. Therefore, $\hat{\theta}_{MAP}$ is replaced with the value that maximizes the maximum likelihood ($\hat{\theta}_{ML}$), a reasonable approximation if the prior does not dominate. Taking the log of Eq. (5) we obtain the log marginal likelihood,

$$M \approx \log P(X|\hat{\theta}_{ML}) + \log \pi(\hat{\theta}) + (d/2)\log(2\pi) + (1/2)\log|\Sigma| - (d/2)\log N \tag{6}$$

The BIC score only retains the terms that vary in N, since asymptotically the terms that are constant in N do not matter. Dropping the constant terms we get,

$$M \approx \log P(X|\hat{\theta}_{ML}) - (d/2)\log N \tag{7}$$

In the model selection problem, we pick the model with the highest BIC score. Frequentist analysis shows that the BIC score is an asymptotically consistent model selection procedure under weak conditions. Note that there is no prior $\pi(\theta)$ in this estimate, so it is clearly not a Bayesian procedure. The BIC score is part of a family of competing penalized likelihood scores that also include the AIC, DIC and TIC scores. These scores differ mostly in the model complexity term $-(d/2)\log N$, where d is the dimensionality of θ, and penalizes models with higher complexity. The AIC score does not have a $\log N$ term and has a fixed dimensionality penalty that allows for more complex models. The goal of the AIC score is to predict the next value and it can be show that it is optimal in the sense of minimizing the Kullback−Leibler (KL) divergence. However, when it comes to model selection, unlike the BIC score, the AIC score is not consistent asymptotically. It should be noted that this method does not provide a particularly good approximation of the marginal likelihood, but is presented as an example of using the Laplace approximation.

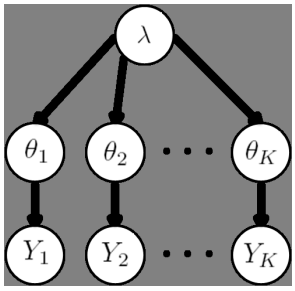

Figure 1: The standard random effects graphical model

5 Full Bayes versus empirical Bayes

Using the standard model from Figure 1, we are now interested in the inference for some function of θ. For generality let us consider $g(\theta)$ to be an arbitrary function of g. Since θ_i is independent of $Y_{j\neq i}$ given λ we have,

$$\begin{aligned} E(g(\theta_i)|Y) &= E_\theta(E(g(\theta_i)|Y_i,\lambda)) \\ Var(g(\theta_i)|Y) &= E_\lambda(Var(g(\theta_i)|Y_i,\lambda)) + Var_\lambda(E(g(\theta_i)|Y_i,\lambda)) \end{aligned} \tag{8}$$

We are interested in comparing the magnitudes of the asymptotic expansions of the expected value and the variance (Bayesian error bars), under empirical Bayes and full Bayes. Every expectation in these two terms involves computing an integral, which we will approximate using the Laplace method.

5.1 Full Bayes

Consider $E[G(\lambda)|Y]$ for some arbitrary G. As an example, G will be the identity function when computing the mean and the square function when computing the second moment.

$$E[G(\lambda)|Y] = \frac{\int G(\lambda)L(\lambda)\pi(\lambda)\ d\lambda}{\int L(\lambda)\pi(\lambda)\ d\lambda} \tag{9}$$

where the likelihood $L(\lambda) = \Pi_{i=1}^K L_i(\lambda)$ and $L_i(\lambda) = P(Y_i|\lambda) = \int P(y_i|\theta_i,\lambda)\ d\theta_i$.

Computing this ratio of integrals is a major application of the Laplace method in Bayesian statistics. The denominator has the form of a likelihood term times a prior term, which is identical to what we have already seen in the marginal likelihood case and can be solved using the standard Laplace approximation. However, the numerator has an extra term. One way to solve this would be to fold in $G(\lambda)$ into $h(\lambda)$ and use the standard approach, but another approach would be to use a more generalized Laplace formulation. Being consistent with the notation used by Kass and Steffey (1989), we define the generalized Laplace integral as $\int b(\lambda)\exp(-Kh(\lambda))\ d\lambda$. By performing a Taylor series expansion in $b(\lambda)$, in addition to the expansions in $h(\lambda)$ and the exponential function, we obtain the following approximation to the first order of the decaying terms:

$$\int b(\lambda)\exp(-Kh(\lambda))\ d\lambda \approx (2\pi/K)^{\frac{m}{2}}\det(D^2h(\hat{\lambda}))^{-\frac{1}{2}}b(\hat{\lambda})\exp(-Kh(\hat{\lambda})\{1+O(1/K)\} \tag{10}$$

Going back to our problem of solving Eq. (9), we define $h(\lambda) = -(1/K)\log L(\lambda)\pi(\lambda)$. Using the above relationship, we can now perform a Laplace expansion on both the numerator and the denominator. Since both terms have the same $h(\lambda)$ term, we can see that most of the terms in this ratio will cancel out.

[1]The BIC is not actually Bayesian, but is derived from a Bayesian point of view

$$\begin{aligned} E[G(\lambda)|y] &= \frac{\int G(\lambda)L(\lambda)\pi(\lambda)\ d\lambda}{\int L(\lambda)\pi(\lambda)\ d\lambda} \\ &= \frac{(2\pi/K)^{\frac{m}{2}}\det(D^2h(\hat{\lambda}))^{-\frac{1}{2}}G(\hat{\lambda})\exp(-Kh(\hat{\lambda})\{1+O(1/K)\}}{(2\pi/K)^{\frac{m}{2}}\det(D^2h(\hat{\lambda}))^{-\frac{1}{2}}\exp(-Kh(\hat{\lambda})\{1+O(1/K)\}} \\ &= G(\hat{\lambda})\frac{(1+O(1/K))}{(1+O(1/K)} \qquad (11) \\ &= G(\hat{\lambda})(1+O(1/K)) \qquad (12) \end{aligned}$$

The transformation from Eq. (11) to Eq. (12) is due to the following:

$$\frac{(1+O(1/K))}{(1+O(1/K))} = \frac{(1+a(1/K)+O(1/K^2))}{(1+b(1/K)+O(1/K^2))} = 1+(a-b)\frac{1}{K}+O\left(\frac{1}{K^2}\right) = 1+O\left(\frac{1}{K}\right) \qquad (13)$$

for some constants a, b.

To calculate the variance we carry out the same computation with G now being equal to G^2. However, asymptotic analysis shows that this answer would not be accurate to $O(1/K)$. To obtain $O(1/K)$ asymptotics for the variance we actually need to compute the second order Laplace expansion. Using the second order expansion as shown in Kass and Steffey (1989) gives us the following result:

$$Var[G(\lambda)|y] = (DG)^T\hat{\Sigma}(DG)\{1+O(1/K)\} \qquad (14)$$

where $\hat{\Sigma} = (-D^2\log(L(\hat{\lambda})\pi(\hat{\lambda})))^{-1}$.

This is often called the delta method. To obtain the expectation of a non-linear function you just evaluate it at the desired point. For the variance you evaluate the second derivative and pre-impulse multiply it by the gradient. The Laplace approximation analysis makes these bounds rigorous in cases where we have asymptotic behavior.

Applying these results to the equality relations from Eqs. (8), we obtain the following. For any function $g(\theta)$,

$$\begin{aligned} E(g(\theta_i)|Y) &= E(g(\theta_i)|Y,\hat{\lambda})(1+O(1/K)) \qquad (15) \\ Var(g(\theta_i)|Y) &= \{Var(g(\theta_i)|Y_i,\hat{\lambda}) + \sum_{jk}\tilde{\sigma}_{jk}\tilde{\delta}_j\tilde{\delta}_k\}(1+O(1/K)) \qquad (16) \end{aligned}$$

where $(\tilde{\Sigma})_{jk} = \tilde{\sigma}_{jk}$ is the matrix of second derivatives and $\tilde{\delta}_j = \frac{\partial}{\partial\lambda_j}(Eg(\lambda_i)|Y_i,\lambda))\big|_{\lambda=\hat{\lambda}}$.

5.2 Empirical Bayes

The empirical Bayesian will not consider a prior on λ and will instead pick a value for λ such as the maximum marginal likelihood $L(\lambda)$. Now we have, $\hat{\lambda}_{EB} = \text{argmax}\ L(\lambda)$. Recall that in full Bayes $\hat{\lambda} = \text{argmax}\ L(\lambda)\pi(\lambda)$. Since taking the logarithm does not change the maximum of a function, $\hat{\lambda}$ can be rewritten as argmax $\log L(\lambda) + \log\pi(\lambda)$. The $\log L(\lambda)$ term grows with the amount of data, but the $\log\pi(\lambda)$ does not, so asymptotically $\hat{\lambda}_{EB} = \hat{\lambda}$ and the empirical Bayes expectation $E(g(\theta_i)|Y)$ is equal to the full Bayes expectation from Eq. (15) within a factor of $O(1/K)$. However, the empirical Bayes variance is not equal to the full Bayes variance within a factor of $O(1/K)$. The empirical Bayes variance is $Var(g(\theta_i)|Y_i,\hat{\lambda})$, but in the full Bayes case from Eq. (16) the variance has an extra term of $\sum_{jk}\tilde{\sigma}_{jk}\tilde{\delta}_j\tilde{\delta}_k$.

Now let us assume that each Y_i is based on n_i data points and analyze the asymptotic behavior of Eq. (8).

$$Var(g(\theta_i)|Y) = E_\lambda(Var(g(\theta_i)|Y_i, \lambda)) + Var_\lambda(E(g(\theta_i)|Y_i, \lambda))$$

Since we will have better error bars for our estimate as n_i increases, $Var(g(\theta_i)|Y_i, \lambda) = O(n_i^{-1})$. This holds when taking the expectation over λ as well. So the first term,

$$E_\lambda(Var(g(\theta_i)|Y_i, \lambda)) = O(n_i^{-1}) \tag{17}$$

Now lets consider the second term $Var_\lambda(E(g(\theta_i)|Y_i, \lambda))$, which is a function of λ. From the Laplace expansion we know this term is equal to a constant plus $O(1/k)$. So we have,

$$Var_\lambda(E(g(\theta_i)|Y_i, \lambda)) = O(K^{-1}) \tag{18}$$

We can see that the two terms have different kinds of asymptotics. The first term depends on the intra-group sample size n_i and the second depends on the number of groups K. If you have a lot of data points per group, the first term will approach zero and the second term will dominate. From Eq. (16) we have,

$$Var(g(\theta_i)|Y) = \{Var(g(\theta_i)|Y_i, \hat{\lambda}) + \sum_{jk} \tilde{\sigma}_{jk}\tilde{\delta}_j\tilde{\delta}_k\}(1 + O(1/K))$$

If K is really large, the second term will go to zero and the first term will dominate. This means that when k is really large, the Bayesian variance is equal to the empirical Bayesian variance within a factor of $O(1/k)$. When n is large the difference between full and empirical Bayesian variance is given by the second term.

In this example we are using Laplace expansion as an analysis tool and not as a procedure. This analysis only tells us how the full Bayesian and empirical Bayesian procedures relate to each other. It can not help us determine which procedure is better.

References

Kass, R. E. and Steffey, D. (1989). Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *Journal of the American Statistical Association*, 84(407):717–726.