Modal and distributional approximations

HwiChang Jeong

November 3, 2020

Seoul National University

Table of Contents

Table of Contents

Finding posterior modes

- The mode is the value that appears most often in a set of data values.
- The posterior mode is often used in statistical practice as a point estimate.
- We summarize the posterior by its mode for computational convenience or as a quick approximation.
- In Bayesian computation, we search for modes as a way to begin mapping the posterior density.
- We discuss algorithms for finding posterior modes.

Conditional maximization

- Simply start somewhere in the target distribution.
- Setting the parameters at rough estimates and then alter one set of components of θ at a time, leaving the other components at their previous value at each step increasing the log posterior density.
- Assuming the posterior density is bounded, the steps will eventually converge to a local mode.

Newton's method

- Iterative approach based on a quadratic Taylor series approximation of the log posterior density.
- It is also acceptable to use an unnormalized posterior density, since uses only the derivatives of $L(\theta) = \log p(\theta|y)$
- The mode-finding algorithm :
  - Choose a staring value, θ^0
  - Set, the new iterate, θ^t, to maximize the quadratic approximation;

$$\theta^t = \theta^{t-1} - [L^{''}(\theta^{t-1})]^{-1} L^{'}(\theta^{t-1}).$$

- The starting value is important ; algorithm is not guaranteed to converge from all starting values, particularly tn regions where $-L^{''}$ is not positive definite.

Table of Contents

Posterior modes on the boundary of parameter space

- The posterior mode is a good point summary of a symmetric posterior distribution.
- If the posterior is asymmetric however, the mode can be a poor point estimate.

Posterior modes on the boundary of parameter space

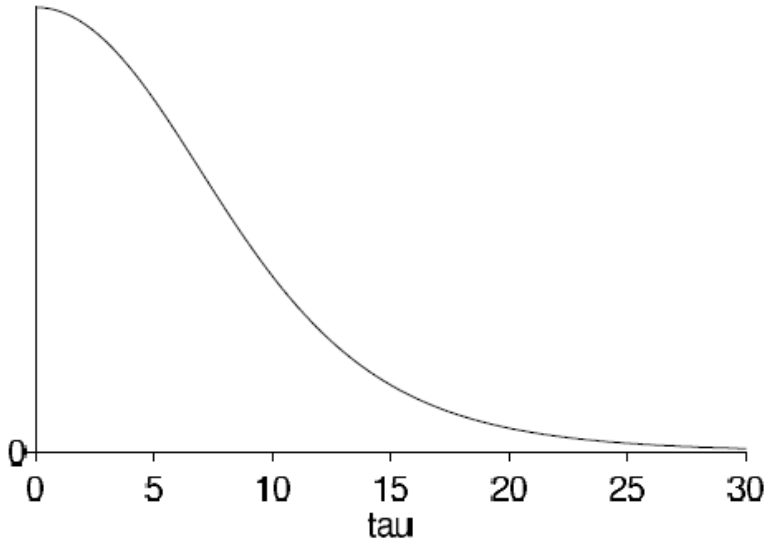


Figure 13.1 *Marginal posterior density, $p(\tau|y)$, for the standard deviation of the population of school effects θ_j in the educational testing example. If we were to choose to summarize this distribution by its mode, we would be in the uncomfortable position of setting $\hat{\tau} = 0$, an estimate on the boundary of parameter space.*

Posterior modes on the boundary of parameter space

- The problem in the above example arise because the mode is taken as a posterior summary.
- If we are planning to summarize the posterior distribution by its mode, it can make sense to choose the prior distribution accordingly.
- We prefer a prior model such as $\tau \sim \mathit{Gamma}(2, \frac{2}{A})$, a gamma distribution with shape 2 and some large scale parameter.
- This density starts at 0 when $\tau = 0$ and then increases linearly from there, eventually curving gently back to zero for large values of τ.

Boundary-avoiding prior distribution for a correlation parameter

- Within each group $j = 1, \cdots, J(= 10)$, we assume a linear model:

$$y_{ij} \sim N(\theta_{j1} + \theta_{j2} x_i + 1), \text{for } i = 1, \cdots n_j(= 5)$$

- The two regression parameters in each group j are modeled as

$$\begin{pmatrix} \theta_{j1} \\ \theta_{j2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \right)$$

- We average over the linear paramters θ and work with the maginal likelihood, which can be computed analytically as

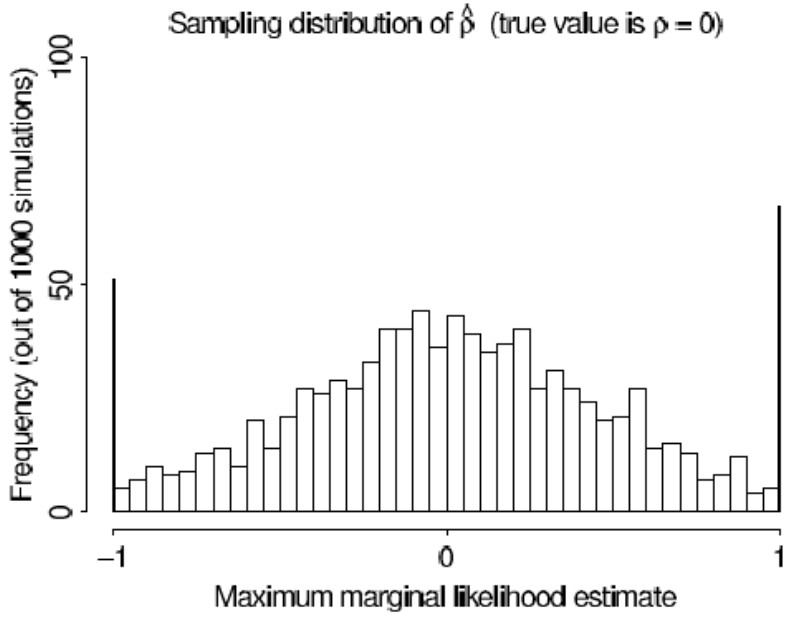
$$p\left(y \mid \tau_1, \tau_2, \rho\right) = \prod_{j=1}^{J} \mathrm{N}\left(\hat{\theta}_j \mid 0, V_j + T\right)$$

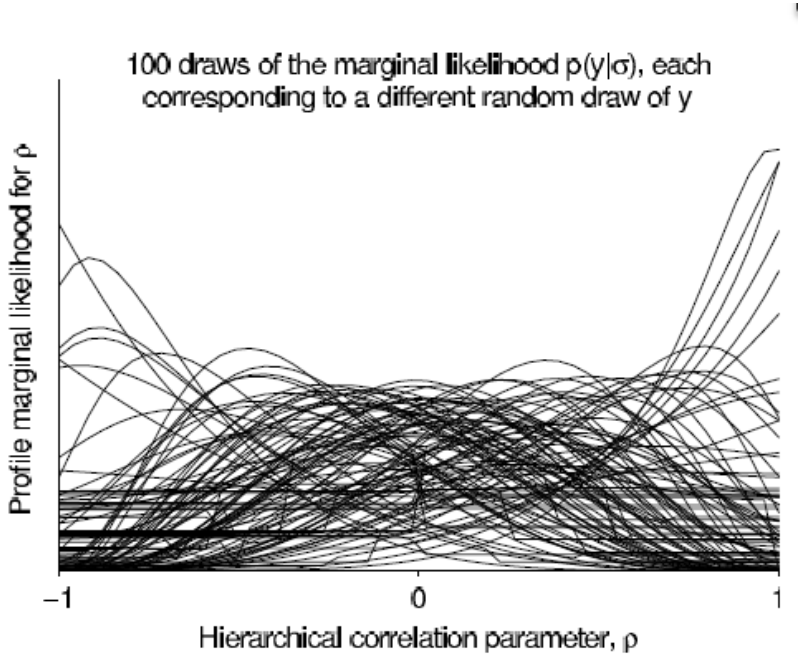
where $\hat{\theta}_j$ and V_j are the LSE and corresponding covariance matrix from regressing y and x for the data group j.

Boundary-avoiding prior distribution for a correlation parameter

- We assume the true values of the variance parameters are $\tau_1 = \tau_2 = 0.5$ and $\rho = 0$
- We simulate data and compute the marginal likelihood $L_{profile}(\rho|y) = \max_{\tau_1,\tau 2} p(y|\tau_1, \tau 2, \rho)$(Uniform prior of ρ).

Posterior modes on the boundary of parameter space





Boundary-avoiding prior distribution for a correlation parameter

- If the plan is to summarize inference by the posterior mode of ρ, we would replace the $U(-1, 1)$ prior distribution with $p(\rho) \propto (1-\rho)(1+\rho)$, which is equivalent to $Beta(2, 2)$ on the transformed parameter $\frac{\rho+1}{2}$
- For a general $d \times d$ covariance matrix we choose the $Wishart(d+3, AI)$ prior density, which is zero but with a positive constant derivative at the boundary.
- In two dimensions, the multivariate model in the limit $A \to \infty$ corresponds to the prior distribution $p(\rho) \propto (1-\rho)(1+\rho)$ as before.

Table of Contents

Fitting multivariate normal densities based on the curvature at the modes

- Once the mode or modes have been found, we can construct an approximation based on the normal distribution.
- For simplicity we first consider the case of a single mode at $\hat{\theta}$, where we fit a normal distribution to the first two derivatives of the log posterior density function at $\hat{\theta}$:

$$p_{\text{normal approx}}(\theta) = N(\theta|\hat{\theta}, V_\theta)$$

where $V_\theta = [-\frac{d^2 \log p(\theta|y)}{d\theta^2}|_{\theta=\hat{\theta}}]^{-1}$

Laplace's method for analytic approximation of integrals

- Instead of approximating just the posterior with normal distribution, we can use Laplace's method to approximate integrals of a smooth function times the posterior $h(\theta)p(\theta|y)$.
- Laplace approximation :

$$\int h(\mathrm{x}) e^{Mf(\mathrm{x})} d\mathrm{x} \approx \left(\frac{2\pi}{M}\right)^{d/2} \frac{h(\mathrm{x}_0)\, e^{Mf(\mathrm{x}_0)}}{\left|-H(f)(\mathrm{x}_0)\right|^{1/2}}$$

  as $M \to \infty$

- When d is dimension of θ, $u(\theta) = \log(h(\theta)p(\theta|y))$,

$$\mathrm{E}(h(\theta) \mid y) \approx \quad h(\theta_0)\, p(\theta_0 \mid y)\, (2\pi)^{d/2} \left|-u''(\theta_0)\right|^{1/2}$$

Mixture approximation for multimodal densities

- Suppose we have found K modes in the posterior density. the target density $p(\theta|y)$ can be approximated by

$$p_{\text{normal approx}}(\theta) \propto \sum_{k=1}^{K} \omega_k \mathrm{N}\left(\theta \mid \hat{\theta}_k, V_{\theta k}\right).$$

- For each k, the mass ω_k of the kth component of the multivariate normal mixture can be estimated by equating the (unnormalized) posterior density $q(\hat{\theta_k}|y)$, to the approximation $p_{\text{normal approx}}(\hat{\theta_k})$ at each of the K modes.
- If the modes are fairly widely separated and the normal approximation is appropriate for each mode, then we obtain $\omega_k = q(\hat{\theta}_k|y)|V_{\theta k}|^{1/2}$

Mixture approximation for multimodal densities

- Normal-mixture approximation :

$$p_{\text{normal approx}}(\theta) \propto \sum_{k=1}^{K} q\left(\hat{\theta}_k \mid y\right) \exp\left(-\frac{1}{2}\left(\theta-\hat{\theta}_k\right)^T V_{\theta k}^{-1}\left(\theta-\hat{\theta}_k\right)\right)$$

- T approximation :

$$p_{t\ \text{approx}}(\theta) \propto \sum_{k=1}^{K} q\left(\hat{\theta}_k \mid y\right)\left(\nu+\left(\theta-\hat{\theta}_k\right)^T V_{\theta k}^{-1}\left(\theta-\hat{\theta}_k\right)\right)^{-(d+\nu)/2}$$

Table of Contents

Finding marginal posterior modes using EM

- In problems with many parameters, normal approximations to the joint distribution are often useless, and the joint mode is typically not helpful.
- It is often useful, however, to base an approximation on marginal posterior mode of a subset of the parameters.
- $\theta = (\gamma, \phi)$ and suppose we are interested in first approximating $p(\phi|y)$.
- After approximating $p(\phi|y)$ as a normal or t or a mixture of these, we may be able to approximate the conditional distribution, $p(\gamma|\phi, y)$ as normal or t or mixture of these with parameters depending on ϕ.

Derivation of the EM

- EM finds the modes of the marginal posterior distribution, $p(\phi|y)$, averaging over the parameters γ.
- We start with the simple identity

$$\log p(\phi|y) = \log p(\gamma, \phi|y) - \log p(\gamma|\phi, y)$$

and take expectation, treating γ as a random variable with the distribution $p(\gamma|\phi^{old}, y)$, where ϕ^{old} is the current guess.

$$\log p(\phi|y) = E_{old}(\log p(\gamma, \phi|y)) - E_{old}(\log p(\gamma|\phi, y))$$

Derivation of the EM

$$\log p(\phi|y) = E_{old}(\log p(\gamma, \phi|y)) - E_{old}(\log p(\gamma|\phi, y))$$

$$E_{old}(\log p(\gamma|\phi, y)) - E_{old}(\log p(\gamma|\phi^{old}, y))$$

$$= E_{old} \log \frac{p(\gamma|\phi, y)}{p(\gamma|\phi^{old}, y)}$$

$$\leq \log E_{old}\left(\frac{p(\gamma|\phi, y)}{p(\gamma|\phi^{old}, y)}\right)$$

$$= \log \int \frac{p(\gamma|\phi, y)}{p(\gamma|\phi^{old}, y)} p(\gamma|\phi^{old}, y) d\gamma$$

$$= log 1 = 0$$

Derivation of the EM

- $E_{old}(\log p(\gamma|\phi, y))$ is maximized at $\phi = \phi^{old}$.
- $E_{old}(\log p(\gamma, \phi|y))$ is called $Q(\phi|\phi^{old})$.
- We increase the $Q(\phi|\phi^{old})$ while not increasing $E_{old}(\log p(\gamma|\phi, y))$ and so the total must increase.
- The EM algorithm can be described algorithmically as follows.
  - Start with a crude parameter estimate, ϕ^0
  - For t=1,2,...;
    - E-step : Determine the expected log posterior density function $Q(\phi|\phi^{t-1})$
    - M-step : $\phi^t = \text{argmax}_\phi\, Q(\phi|\phi^{t-1})$

Example

- Suppose, $y_1, \cdots, y_n$ *iid* $N(\mu, \sigma^2)$.
- We assume, $\mu \sim N(\mu_0, \tau_0^2)$ prior distribution on μ and the standard noninformative uniform prior distribution on $\log \sigma$.
- We use EM algorithm to find the marginal posterior mode of μ.
- Joint log posterior :

$$\log p(\mu, \sigma \mid y) = -\frac{1}{2\tau_0^2}(\mu - \mu_0)^2 - (n+1)\log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 + C$$

Example

- E-step :

$$\mathrm{E}_{\text{old}}\ \log p(\mu, \sigma \mid y) = -\frac{1}{2\tau_0^2}(\mu - \mu_0)^2 - (n+1)\mathrm{E}_{\text{old}}\,(\log \sigma) - \frac{1}{2}\mathrm{E}_{\text{old}}\left(\frac{1}{\sigma^2}\right)\sum_{i=1}^{n}(y_i - \mu)^2 + \text{ constant.}$$

- The posterior distribution of σ^2 given μ is scaled inverse-χ^2

$$\mathrm{E}_{\text{old}}\left(\frac{1}{\sigma^2}\right) = \mathrm{E}\left(\frac{1}{\sigma^2} \mid \mu^{\text{old}}, y\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu^{\text{old}}\right)^2\right)^{-1}$$

$$\mathrm{E}_{\text{old}} \log p(\mu, \sigma \mid y) = -\frac{1}{2\tau_0^2}(\mu - \mu_0)^2 - \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu^{\text{old}}\right)^2\right)^{-1}\sum_{i=1}^{n}(y_i - \mu)^2 + \text{const.}$$

Example

- M-step : Find the μ that maximizes the above expression.
- Marginal posterior distribution of μ has the form of a normal distribution, M-step is achieved by the mode of the density

$$\mu^{\text{new}} = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu^{\text{old}}\right)^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu^{\text{old}}\right)^2}}.$$

Table of Contents

Approximating conditional and marginal posterior densities

- The normal, t, and other analytically convenient distributions can be poor approximations to a joint posterior distribution.
- However, we can partition the parameter vector as $\theta = (\gamma, \phi)$, in such a way that an analytic approximation works well for the conditional posterior density, $p(\gamma|\phi, y)$
- The mode-finding techniques and normal approximation can be applied directly to the marginal posterior density if the marginal distribution can be obtained analytically.
- If not, the EM algorithm may allow us to find the mode of the marginal posterior density and construct an approximation.

Approximating conditional and marginal posterior densities

- On occasion it is not possible to construct an approximation to $p(\phi|y)$ using any of these methods, if we have an analytic approximation to the conditional posterior density, $p(\gamma|\phi, y)$, we may derive an approximation.

$$p_{\text{approx}}(\phi \mid y) = \frac{p(\gamma, \phi \mid y)}{p_{\text{approx}}(\gamma \mid \phi, y)}$$

- We must specify a value γ(possibly as a function of ϕ) since the left side does not involve γ at all.

Improving an approximation using importance sampling.

- We can improve the approximation with importance sampling, using draws of γ from each value of ϕ at which the approximation is computed.

$$
\begin{aligned}
p(\phi \mid y) &= \int p(\gamma, \phi \mid y) d\gamma \\
&= \int \frac{p(\gamma, \phi \mid y)}{p_{\text{approx}}\left(\gamma \mid \phi, y\right)} p_{\text{approx}}\left(\gamma \mid \phi, y\right) d\gamma \\
&= \mathrm{E}_{\text{approx}}\left(\frac{p(\gamma, \phi \mid y)}{p_{\text{approx}}\left(\gamma \mid \phi, y\right)}\right)
\end{aligned}
$$

E_{approx} averages over γ using the conditional posterior distribution, $p_{approx}(\gamma|\phi, y)$.

Table of Contents

Variational inference

- EM proceeds by alternately evaluating conditional expectations of the log density and using these to maximize a function of a set of hyperparameters
- In variational bayes, the iterations lead to a closed-form approximation that is the closest fit to the posterior distribution within some specified class of functions.
- A parametric approximation $g(\theta)$ is constructed iteratively using an expectation procedure that, as we shall show, has the effect of minimizing the Kullback-Leibler divergence from the target posterior distribution $p(\theta|y)$,

$$\mathrm{KL}(g\|p) = -\mathrm{E}_g\left(\log\left(\frac{p(\theta \mid y)}{g(\theta)}\right)\right) = -\int \log\left(\frac{p(\theta \mid y)}{g(\theta)}\right) g(\theta) d\theta$$

Variational inference

- We shall use the notation ϕ for the hyperparameters of the variational approximation. Thus we write $g(\theta)$ as $g(\theta|\phi)$.
- The algorithm proceeds by staring with some guess of ϕ and then iteratively updating it in a way that is mathematically guaranteed to decrease the Kullback-Leibler divergence at each step.
- It can make sense to check the results by running the algorithm several times from different starting points.

The class of approximate distributions

- There are various ways of defining the class of distributions for $g(\theta|\phi)$.
- A standard approach is to constrain th components of θ to be independent:

$$g(\theta \mid \phi) = \prod_{j=1}^{J} g_j(\theta_j \mid \phi_j)$$

  for a J-dimensional parameter θ.
- For each j, we examine the expectation of the log posterior density, $\log p(\theta|y)$, considering it as a function of θ_j, averaging over the distributions g_{-j} that represent the other $J-1$ dimensions of θ.
- We do not need to evaluate the expectation; we merely need to figure out its mathematical form as a function of θ_j.

The Variational Bayes algorithm

- Once the classes of approximating distributions $g_j(\theta_j|\phi_j)$ have been identified, the computation begins with guesses of all the hyperparameters ϕ.
- We then cycle through the distributions g_j, in each of these steps updating the hyperparameters ϕ_j so that $\log g_j(\theta_j|\phi_j)$ is set to $E_{g_{-j}}(\log p(\theta|y)) = \int \log p(\theta|y) g_{-j}(\theta_{-j}|\phi_{-j}) d\theta_{-j}$
- The steps of variational Bayes decrease $KL(g||p)$ and thus gradually bring the approximating distribution $g(\theta)$ closer to the target posterior distribution $p(\theta|y)$.

Example

- Suppose full vector of parameters θ has 10 dimesions, corresponding to $\alpha_1, \alpha_2, \cdots, \alpha_8, \mu, \tau$, and the log posterior density is

$$\log p(\theta \mid y) = -\frac{1}{2}\sum_{j=1}^{8}\frac{(y_j - \alpha_j)^2}{\sigma_j^2} - 8\log\tau - \frac{1}{2}\frac{1}{\tau^2}\sum_{j=1}^{8}(\alpha_j - \mu)^2 + \text{const.}$$

- Independent densities:

$$g(\theta) = g(\alpha_1, \ldots, \alpha_8, \mu, \tau) = g(\alpha_1)\cdots g(\alpha_8)\, g(\mu) g(\tau).$$

Example

- $$g\left(\alpha_j\right) = \mathrm{N}\left(\alpha_j \mid \frac{\frac{1}{\sigma_j^2} y_j + \mathrm{E}\left(\frac{1}{\tau^2}\right) \mathrm{E}(\mu)}{\frac{1}{\sigma_j^2} + \mathrm{E}\left(\frac{1}{\tau^2}\right)}, \frac{1}{\frac{1}{\sigma_j^2} + \mathrm{E}\left(\frac{1}{\tau^2}\right)}\right)$$

$$g(\mu) = \mathrm{N}\left(\mu \mid \frac{1}{8} \sum_{j=1}^{8} \mathrm{E}\left(\alpha_j\right), \frac{1}{8} \frac{1}{\mathrm{E}\left(\frac{1}{\tau^2}\right)}\right)$$

$$g\left(\tau^2\right) = \text{Inv-}\chi^2\left(\tau^2 \mid 7, \frac{1}{7} \sum_{j=1}^{8} \mathrm{E}\left(\left(\alpha_j - \mu\right)^2\right)\right)$$

- Rewrite above factors,

$$g\left(\alpha_j\right) = \mathrm{N}\left(\alpha_j \mid M_{\alpha_j}, S_{\alpha_j}^2\right), \text{ for } j = 1, \ldots, 8$$

$$g(\mu) = \mathrm{N}\left(\mu \mid M_\mu, S_\mu^2\right)$$

$$g\left(\tau^2\right) = \text{Inv} - \chi^2\left(\tau^2 \mid 7, M_\tau^2\right)$$

Example

- For simplicity, we draw the unbounded parameters $M_{\alpha_1}, \cdots, M_{\alpha_8}, M_\mu$ from independent $N(0,1)$ and draw the bounded parameters, $S_{\alpha_1}, \cdots, S_{\alpha_8}, S_\mu$ from independent $U(0,1)$.
- We iterate through α, μ, τ at each iteration updating distributions. Then we turn around and label the newly computed means and standard deviations as the updated $M's$ and $S's$.
- Difference with EM is that VI is distributions rather than point estimates.

Variational Bayes followed by importance sampling

- Variational methods are commonly used as an approximate method when simulation-based full Bayes is too computationally expensive, as with very large models or datasets.
- It might make sense to use the variational estimate as a starting point for a stochastic algorithm leading to a better approximation to the target distribution.
- Compute S simulation draws, θ^s from g and for each compute the importance weight $\frac{p(\theta^s|y)}{g(\theta^s)}$.
- As usual we only need theses weights up to an arbitrary multiplicative constant, thus it would be fine to use unnormalized densities.

Variational Bayes followed particle filtering

- The distribution of importance ratios can have long tails, leading to unstable averages.
- We would recommend without replacement sampling.
- A more general approach would be particle filtering, using draws from the variational bayes as a starting point and then moving through the target density using Metropolis or Hamiltonian Monte Carlo and splitting and removing points as appropriate.

Table of Contents

Expectation propagation

- Expectation propagation is another deterministic iterative algorithm in which the posterior distribution $p(\theta|y)$ is approximated by a best-fit distribution from some specified parametric family.
- We first describe the algorithm in general and then go through the steps of applying to logistic regression.
- The target distribution $p(\theta|y)$, which we write as $f(\theta)$, suppressing the dependence on y which is not directly relevant for these computations.
- We assume :

$$f(\theta) = \prod_{i=0}^{n} f_i(\theta)$$

As with many Bayesian computations, $f_i's$ are the unnormalized density functions.

Expectation propagation

- Expectation propagation can be expressed more generally, but it is think of $f_0(\theta)$ as prior density and each $f_i(\theta)$ as the likelihood for one data point.
- A key difference between VI and EP is that variational inference is typically based on a separation of g into factors for each 'parameter', whereas expectation propagation factorizes g based on a partition of the 'data'.

Exponential families, sufficient statistics, and natural parameters

- The approximating distribution $g(\theta)$ should be in the exponential family.
- This means that the density can be written as a normalizing function times the exponential of a linear function of 'sufficient statistics' of θ.
- The coefficients of the sufficient statistics inside the exponential are called the natural parameters of the model.

The expectation propagation algorithm

- Define the (unnormalized)

$$\text{cavity distribution: } g_{-i}(\theta) \propto \frac{g(\theta)}{g_i(\theta)}$$

$$\text{tilted distribution: } g_{-i}(\theta) f_i(\theta)$$

- We construct an approximation to the tilted distribution, using a moment-matching approach.
- This approach is the updated $g(\theta)$. Then back out the updated approximating factor, $g_i(\theta) = g(\theta)/g_{-i}(\theta)$.
- The result is that we have a new $g_i(\theta)$ which approximates $f_i(\theta)$.
- Moment matching : Setting the expectation of the sufficient statistics of g to the corresponding expectations of θ in $g_{-i}(\theta) f_i(\theta)$.

The expectation propagation algorithm

- For example, if $g(\theta)$ has the form $N(\theta|\mu, \Sigma)$, then tn the moment-matching step we set
  $\mu = \mathrm{E}_{\mathrm{tilted}i}(\theta) = \int \theta g_{-i}(\theta) f_i(\theta) d\theta$ and
  $\Sigma = \mathrm{var}_{\mathrm{tilted}i}(\theta) = \int (\theta - \mu)(\theta - \mu)^T g_{-i}(\theta) f_i(\theta) d\theta$.
- In practical implementations of expectation propagation, these integrals can be done in closed form or via a transformation that reduces the problem to a low-dimensional integral.
- If g is updated after each moment-matching step, the algorithm is called sequential EP, whereas if g is updated only after all tilted moments have been computed the algorithm is called parallel EP.
- Parallel EP is typically much faster as it requires less frequent updates of the higher-dimensional function of g.

The expectation propagation algorithm

- Moment matching corresponds to minimizing the KL-divergence from the tilted distribution to the new approximated marginal distribution, but the iterative matching of the marginals does not guarantee that the KL from the full posterior distribution to the overall approximation is minimized.
- There is no guarantee convergence for EP, but the algorithm has been used successfully in many applications.

Table of Contents

Integrated nested Laplace approximation(INLA)

- INLA is another form of posterior approximation involves partitioning the parameters into a large set γ conditional on a smaller set of hyperparameters ϕ.
- Construct a joint gaussian approximation for $p(\gamma|\phi, y)$ and approximate $p(\phi|y)$ and $p(\gamma_i|\phi, y)$.
- Approximations to $p(\gamma_i|y)$ are obtained by numerically integrating over the low dimensional $p_{approx}(\phi|y)$.
- INLA works best when there are not many hyperparameters in the model.

Approximate Bayesian computation(ABC)

- ABC is applied to set of statistical procedures based on drawing parameters θ from an initial or approximate distribution, then sampling replicated data $y^{rep}|\theta$ from the model, and then accepting or rejecting the sample based on the closeness of y^{rep} to the observed data y.
- The attraction of ABC is that it does not require computation of the likelihood function, only the ability to simulate $y^{rep}|\theta$ from the data distribution.

Approximate Bayesian computation(ABC)

- ABC has the form of simple rejection sampling:
  - Draw θ from the prior distribution $p(\theta)$ and then y^{rep} from the data distribution, $p(y^{rep}|\theta)$, thus obtaining a single draw of y^{rep} from its marginal distribution.
  - Compute a discrepancy measure $d(y^{rep}, y)$, where d is defined so that it is zero if y and y^{rep} are identical and is larger the more different they are in some relevant dimensions.
  - Accept θ if $d(y^{rep}, y) < \epsilon$ for some preset threshold ϵ, otherwise reject.
- ABC involves three challenges.
  - One needs to define a discrepancy measure d
  - ϵ needs to be set small enough that the data provide information, but not so small that all the simulations get rejected.
  - If the prior distribution is broad enough, the rejection rate can be unacceptably high even if the discrepancy measure and threshold have been chosen well.

Approximate Bayesian computation(ABC)

- ABC has the form of simple rejection sampling:
  - Draw θ from the prior distribution $p(\theta)$ and then y^{rep} from the data distribution, $p(y^{rep}|\theta)$, thus obtaining a single draw of y^{rep} from its marginal distribution.
  - Compute a discrepancy measure $d(y^{rep}, y)$, where d is defined so that it is zero if y and y^{rep} are identical and is larger the more different they are in some relevant dimensions.
  - Accept θ if $d(y^{rep}, y) < \epsilon$ for some preset threshold ϵ, otherwise reject.
- ABC involves three challenges.
  - One needs to define a discrepancy measure d
  - ϵ needs to be set small enough that the data provide information, but not so small that all the simulations get rejected.
  - If the prior distribution is broad enough, the rejection rate can be unacceptably high even if the discrepancy measure and threshold have been chosen well.

Table of Contents

Unknown normalizing factors in the likelihood

- A new problem arises when the sampling density $p(y|\theta)$ has an unknown normalizing factor that depends on θ.
- Such models often arise in problems that are specified conditionally, such as in spatial statistics.
- In general we use the following notation:

$$p(y|\theta) = \frac{1}{z(\theta)} q(y|\theta),$$

and

$$z(\theta) = \int q(y|\theta) dy$$

- $z(\theta)$ is called the normalizing factor of the family of distributions, we can no longer call it a constant.
- Then the posterior density :

$$p(\theta|y) \propto p(\theta) \frac{1}{z(\theta)} q(y|\theta).$$

Posterior computations involving an unknown normalizing factor

- Obtain an analytic estimates of $z(\theta)$ using some approximate method(Laplace method).
- Construct an approximation to the posterior distributions.
- If θ is only one or two-dimensional it may be reasonable to compute $z(\theta)$ over a finite grid and interpolate to obtain an estimate of $z(\theta)$ as a function of θ.

Computing the normalizing factor

- The normalizing factor can be computed, for each value of θ, using some of the numerical integration approaches.
- The importance sampling estimated is based on the identity

$$z(\theta) = \int \frac{q(y \mid \theta)}{g(y)} g(y) dy = \mathrm{E}_g \left(\frac{q(y \mid \theta)}{g(y)} \right).$$

  where E_g averages over y under the approximation density $g(y)$.
- The estimate of $z(\theta)$ is $\frac{1}{S} \sum_{s=1}^{S} q\left(y^s \mid \theta\right) / g\left(y^s\right)$, based on simulations y^s from $g(y)$.