Bayesian Large Sample Theory and Posterior Approximation

PhD course in Statistical Inference

Mattias Villani

Department of Statistics
Stockholm University
and
Department of Computer and Information Science
Linköping University

- **Large sample theory**
- **Variational inference**

- Response y conditional on a set of covariates $\mathbf{x}$ belongs to the exponential family with dispersion parameter κ

$$p(y|\theta, \kappa, \mathbf{x}) = h(y, \kappa) \exp\left(\frac{\phi(\theta)^T \mathbf{t}(y) - A(\theta)}{d(\kappa)}\right)$$

- The conditional mean of $\mu = E(y|\mathbf{x})$ is a function of a **linear predictor** $\eta = \mathbf{x}^T\beta$ through a **link function**

$$g(\mu) = \mathbf{x}^T\beta$$

- Example: Poisson regression, where $y|\mathbf{x} \sim \mathrm{Pois}(\exp(\mathbf{x}^T\beta))$ is a GLM with log-link: $\log \mu = \mathbf{x}^T\beta$.
- Logistic regression: $y_i|\mathbf{x}_i \sim \mathrm{Bern}(\theta_i)$, where $\theta_i = \frac{\exp(\mathbf{x}_i^T\beta)}{1+\exp(\mathbf{x}_i^T\beta)}$.
- Prior $\beta \sim N(0, \tau^2 I)$.
- Posterior is typically non-standard. What to do?

- **Taylor expansion of log-posterior** around mode $\theta = \tilde{\theta}$:

$$\ln p(\theta|\mathbf{y}) = \ln p(\tilde{\theta}|\mathbf{y}) + \frac{\partial \ln p(\theta|\mathbf{y})}{\partial\theta}\Big|_{\theta=\tilde{\theta}}(\theta - \tilde{\theta}) + \frac{1}{2!}\frac{\partial^2 \ln p(\theta|\mathbf{y})}{\partial\theta^2}\Big|_{\theta=\tilde{\theta}}(\theta - \tilde{\theta})^2 + \ldots$$

- From the definition of the posterior mode:

$$\frac{\partial \ln p(\theta|\mathbf{y})}{\partial\theta}\Big|_{\theta=\tilde{\theta}} = 0$$

- So, in **large samples** (higher order terms negligible):

$$p(\theta|\mathbf{y}) \approx p(\tilde{\theta}|\mathbf{y}) \exp\left(-\frac{1}{2}J_{\mathbf{y}}(\tilde{\theta})(\theta - \tilde{\theta})^2\right)$$

  where $J_{\mathbf{y}}(\tilde{\theta}) = -\frac{\partial^2 \ln p(\theta|\mathbf{y})}{\partial\theta^2}\Big|_{\theta=\tilde{\theta}}$ is the **observed information**.
- **Approximate normal posterior** in large samples.

$$\theta|\mathbf{y} \overset{approx}{\sim} N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$$

Example: gamma posterior

- **Poisson model**: $\theta | y_1, \ldots, y_n \sim Gamma(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$

$$\log p(\theta | y_1, \ldots, y_n) \propto (\alpha + \sum_{i=1}^{n} y_i - 1) \log \theta - \theta(\beta + n)$$

- First derivative of log density

$$\frac{\partial \ln p(\theta|\mathbf{y})}{\partial \theta} = \frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\theta} - (\beta + n)$$

$$\tilde{\theta} = \frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\beta + n}$$

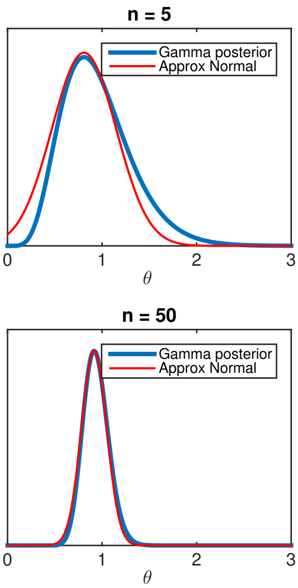
- Second derivative at mode $\tilde{\theta}$

$$\frac{\partial^2 \ln p(\theta|\mathbf{y})}{\partial \theta^2}\Big|_{\theta=\tilde{\theta}} = -\frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\left(\frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\beta + n}\right)^2} = -\frac{(\beta + n)^2}{\alpha + \sum_{i=1}^{n} y_i - 1}$$

- **Normal approximation**

$$N\left[\frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\beta + n}, \frac{\alpha + \sum_{i=1}^{n} y_i - 1}{(\beta + n)^2}\right]$$

Example: gamma posterior



- $\theta|\mathbf{y} \overset{approx}{\sim} N\left[\tilde{\theta}, J_y^{-1}(\tilde{\theta})\right]$ works also when θ is a vector.
- How to compute $\tilde{\theta}$ and $J_{\mathbf{y}}(\tilde{\theta})$?
- Standard **optimization routines** may be used. (`optim.r`).
  - **Input**: expression proportional to $\log p(\theta|\mathbf{y})$. Initial values.
  - **Output**: $\log p(\tilde{\theta}|\mathbf{y})$, $\tilde{\theta}$ and Hessian matrix $(-J_{\mathbf{y}}(\tilde{\theta}))$.
- **Re-parametrization** may improve normal approximation. [Don't forget the **Jacobian**!]
  - If $\theta \geq 0$ use $\phi = \log(\theta)$.
  - If $0 \leq \theta \leq 1$, use $\phi = \ln[\theta/(1-\theta)]$.
- **Heavy tailed approximation**: $\theta|\mathbf{y} \overset{approx}{\sim} t_v\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$ for suitable degrees of freedom v.

Reparametrization - Gamma posterior

- Poisson model. Reparameterize to $\phi = \log(\theta)$.
- Change-of-variables formula from a basic probability course

$$\log p(\phi|y_1, ..., y_n) \propto (\alpha + \textstyle\sum_{i=1}^n y_i - 1)\phi - \exp(\phi)(\beta + n) + \phi$$

- Taking first and second derivatives and evaluating at $\tilde{\phi}$ gives

$$\tilde{\phi} = \log\left(\frac{\alpha + \sum_{i=1}^n y_i}{\beta + n}\right) \text{ and } \frac{\partial^2 \ln p(\phi|y)}{\partial\phi^2}|_{\phi=\tilde{\phi}} = \alpha + \textstyle\sum_{i=1}^n y_i$$

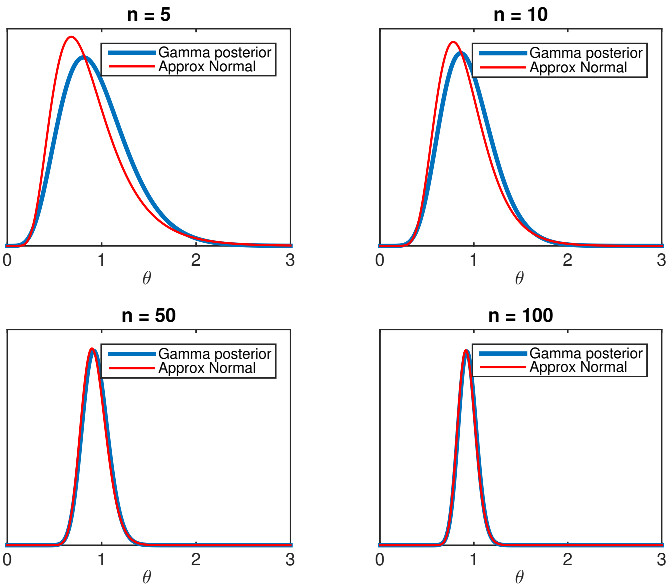
- So, the normal approximation for $p(\phi|y_1, ...y_n)$ is

$$\phi = \log(\theta) \sim N\left[\log\left(\frac{\alpha + \sum_{i=1}^n y_i}{\beta + n}\right), \frac{1}{\alpha + \sum_{i=1}^n y_i}\right]$$

which means that $p(\theta|y_1, ...y_n)$ is log-normal:

$$\theta|\mathbf{y} \sim LN\left[\log\left(\frac{\alpha + \sum_{i=1}^n y_i}{\beta + n}\right), \frac{1}{\alpha + \sum_{i=1}^n y_i}\right]$$

Reparametrization - Gamma posterior



Normal approximation of posterior

- Even if the posterior of θ is approx normal, **interesting functions** of $g(\theta)$ may not be (e.g. predictions).

- But approximate posterior of $g(\theta)$ can be obtained by **simulating** from $N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$.

- Posterior of **Gini coefficient**
  - Model: $x_1, ..., x_n | \mu, \sigma^2 \sim LN(\mu, \sigma^2)$.
  - Let $\phi = \log(\sigma^2)$. And $\theta = (\mu, \phi)$.
  - Joint posterior $p(\mu, \phi)$ may be approximately normal: $\theta | \mathbf{y} \overset{approx}{\sim} N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$.
  - Simulate $\theta^{(1)}, ..., \theta^{(N)}$ from $N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$. Compute $\sigma^{(1)}, ..., \sigma^{(N)}$.
  - Compute $G^{(i)} = 2\Phi\left(\sigma^{(i)}/\sqrt{2}\right)$ for $i = 1, ..., N$.

- Let $\theta = (\theta_1, ..., \theta_p)$. Approximate the posterior $p(\theta|y)$ with a (simpler) distribution $q(\theta)$.
- Before: **Normal approximation** from optimation: $q(\theta) = N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$.
- **Mean field Variational Bayes** (**VB**)

$$q(\theta) = \prod_{i=1}^{p} q_i(\theta_i)$$

- Find the $q(\theta)$ that **minimizes the Kullback-Leibler distance** between the true posterior p and the approximation q:

$$KL(q,p) = \int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} d\theta = E_q\left[\ln \frac{q(\theta)}{p(\theta|y)}\right].$$

- **Mean field VB** is based on factorized approximation:

$$q(\theta) = \prod_{i=1}^{p} q_i(\theta_i)$$

- **No specific functional forms** are assumed for the $q_i(\theta)$.

- **Optimal densities** can be shown to satisfy:

$$q_i(\theta) \propto \exp\left(E_{-\theta_i} \ln p(\mathbf{y}, \theta)\right)$$

where $E_{-\theta_i}(\cdot)$ is the expectation with respect to $\prod_{i \neq j} q_j(\theta_j)$.

- **Structured mean field approximation**. Group subset of parameters in tractable blocks. Similar to Gibbs sampling.

- Initialize: $q_2^*(\theta_2), \ldots, q_M^*(\theta_p)$
- Repeat until convergence:
  - $q_1^*(\theta_1) \leftarrow \frac{\exp\left[E_{-\theta_1} \ln p(\mathbf{y},\theta)\right]}{\int \exp\left[E_{-\theta_1} \ln p(\mathbf{y},\theta)\right] d\theta_1}$
  - $\vdots$
  - $q_p^*(\theta_p) \leftarrow \frac{\exp\left[E_{-\theta_p} \ln p(\mathbf{y},\theta)\right]}{\int \exp\left[E_{-\theta_p} \ln p(\mathbf{y},\theta)\right] d\theta_p}$
- Note: no assumptions about parametric form of the $q_i(\theta)$.
- Optimal $q_i(\theta)$ often **turn out** to be parametric (normal etc).
- Just update hyperparameters in the optimal densities.

- **Model**: $X_i|\theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$.
- **Prior**: $\theta \sim N(\mu_0, \tau_0^2)$ **independent** of $\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2)$.
- **Mean-field approximation**: $q(\theta, \sigma^2) = q_\theta(\theta) \cdot q_{\sigma^2}(\sigma^2)$.
- Optimal densities

$$q_\theta^*(\theta) \propto \exp\left[E_{q(\sigma^2)} \ln p(\theta, \sigma^2, \mathbf{x})\right]$$
$$q_{\sigma^2}^*(\sigma^2) \propto \exp\left[E_{q(\theta)} \ln p(\theta, \sigma^2, \mathbf{x})\right]$$

- **Variational density for σ^2**

$$\sigma^2 \sim Inv-\chi^2\left(\tilde{\nu}_n, \tilde{\sigma}_n^2\right)$$

where $\tilde{\nu}_n = \nu_0 + n$ and $\tilde{\sigma}_n = \frac{\nu_0\sigma_0^2 + \sum_{i=1}^n (x_i - \tilde{\mu}_n)^2 + n\cdot\tilde{\tau}_n^2}{\nu_0 + n}$

- **Variational density for θ**

$$\theta \sim N\left(\tilde{\mu}_n, \tilde{\tau}_n^2\right)$$

where

$$\tilde{\tau}_n^2 = \frac{1}{\frac{n}{\tilde{\sigma}_n^2} + \frac{1}{\tau_0^2}}$$

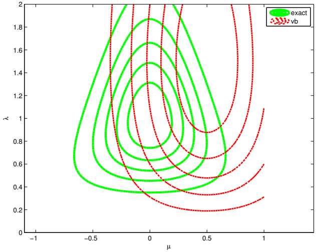
$$\tilde{\mu}_n = \tilde{w}\bar{x} + (1-\tilde{w})\mu_0,$$

where

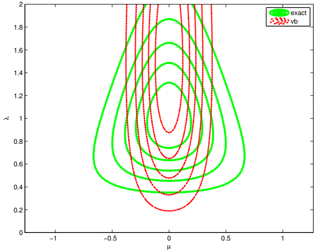
$$\tilde{w} = \frac{\frac{n}{\tilde{\sigma}_n^2}}{\frac{n}{\tilde{\sigma}_n^2} + \frac{1}{\tau_0^2}}$$

Normal example from Murphy ($\lambda = 1/\sigma^2$)

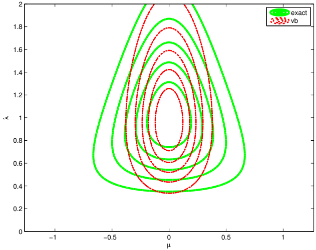
Initial values



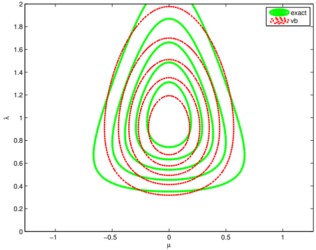
After updating q_μ



After updating q_{σ^2}



At convergence



- **Model**:

$$\Pr(y_i = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_i^T \beta)$$

- **Prior**: $\beta \sim N(0, \Sigma_\beta)$. For example: $\Sigma_\beta = \tau^2 I$.
- **Latent variable formulation** with $u = (u_1, ..., u_n)'$

$$\mathbf{u}|\beta \sim N(\mathbf{X}\beta, 1)$$

and

$$y_i = \begin{cases} 0 & \text{if } u_i \leq 0 \\ 1 & \text{if } u_i > 0 \end{cases}$$

- Factorized **variational approximation**

$$q(\mathbf{u}, \beta) = q_{\mathbf{u}}(\mathbf{u}) q_\beta(\beta)$$

- **VB posterior**

$$\beta \sim N\left(\tilde{\mu}_\beta, \left(\mathbf{X}^T\mathbf{X} + \Sigma_\beta^{-1}\right)^{-1}\right)$$

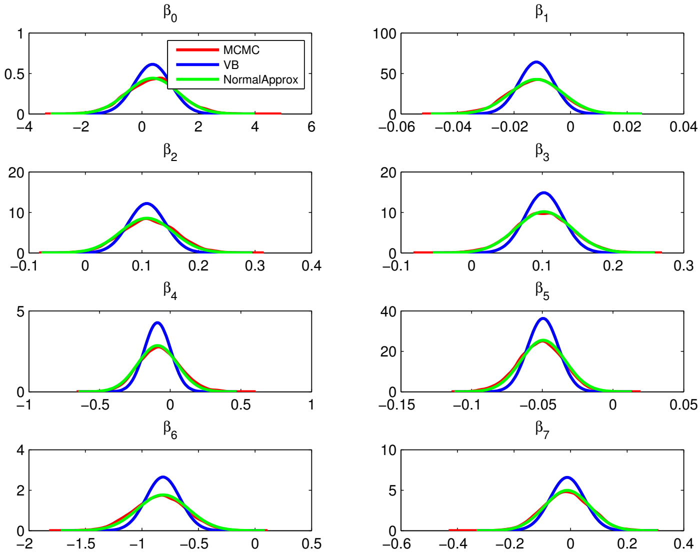
where

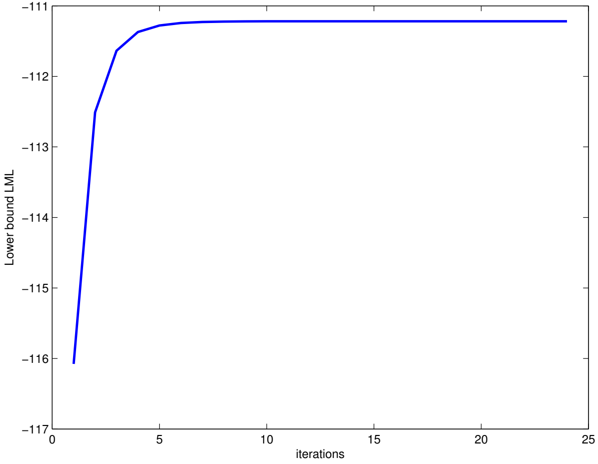
$$\tilde{\mu}_\beta = \left(\mathbf{X}^T\mathbf{X} + \Sigma_\beta^{-1}\right)^{-1}\mathbf{X}^T\tilde{\mu}_\mathbf{u}$$

and

$$\tilde{\mu}_\mathbf{u} = \mathbf{X}\tilde{\mu}_\beta + \frac{\phi\left(\mathbf{X}\tilde{\mu}_\beta\right)}{\Phi\left(\mathbf{X}\tilde{\mu}_\beta\right)^\mathbf{y}\left[\Phi\left(\mathbf{X}\tilde{\mu}_\beta\right) - \mathbf{1}_n\right]^{\mathbf{1}_n - \mathbf{y}}}.$$

Probit example (n=200 observations)




−111
−112
−113
−114
−115
−116
−117
Lower bound LML
0
5
10
15
20
25
iterations