Explaining Variational Approximations

Ricardo Cortez & Ben Graf

28 Apr 2021

Outline

- Introduction
- Density Transform Approach
- Tangent Transform Approach
- Conclusion

Introduction

NEED STUFF HERE! VERY IMPORTANT!

Density Transform Approach

- Approximates intractable posterior densities with better known and easier to deal with densities
- Two main types of restrictions for the q density:
  - Product Density Transforms (non-parametric)
  - Parametric Density Transforms (parametric)
- Guided by Kullback-Leibler(K-L) divergence
  - provides a lower bound which can be maximized in order to minimize the K-L divergence between q and p(.|y)

Kullback-Leibler Divergence

- Let q be an arbitrary density function over Θ. Then the log of the marginal likelihood satisfies:

$$log(p(\boldsymbol{y})) \geq \int q(\theta) log\{\frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\} d\boldsymbol{\theta}$$

- for all densities q, if and only if

$$q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{y})$$

\- it follows immediately that

$$p(\boldsymbol{y}) \geq \underline{p}(\boldsymbol{y}; q)$$

\- where $\underline{p}(\boldsymbol{y}; q)$ is the lower bound on the marginal likelihood

Product Density Transforms

- Suppose q is subject to the product restriction from the previous slide
- It can be shown that the optimal densities satisfy

$$q_i^*(\boldsymbol{\theta}_i) \propto \exp\left\{E_{-\boldsymbol{\theta}_i} \log p(\mathbf{y}, \boldsymbol{\theta})\right\}, \quad 1 \leq i \leq M$$

  where $E_{-\boldsymbol{\theta}_i}$ denotes the expectation of the density with q_i removed
- This leads to the algorithm on the next slide to solve for the q_i^*
- Notes:
  - Can show that convergence to at least local optima guaranteed
  - If conjugate priors used, then the q_i^* updates reduce to updating parameters in a density family
  - Common to monitor convergence using $\log \underline{p}(\boldsymbol{y}; q)$

Product Density Transforms

Algorithm 1 Iterative scheme for obtaining the optimal densities under product density restriction. The updates are based on the solutions given on the previous slide.

Initialize: $q_2^*(\boldsymbol{\theta}_2), \ldots, q_M^*(\boldsymbol{\theta}_M)$.

Cycle:

$$q_1^*(\boldsymbol{\theta}_1) \leftarrow \frac{\exp\left\{E_{-\boldsymbol{\theta}_1} \log p(\mathbf{y},\boldsymbol{\theta})\right\}}{\int \exp\left\{E_{-\boldsymbol{\theta}_1} \log p(\mathbf{y},\boldsymbol{\theta})\right\} d\boldsymbol{\theta}_1},$$

$$\vdots$$

$$q_M^*(\boldsymbol{\theta}_M) \leftarrow \frac{\exp\left\{E_{-\boldsymbol{\theta}_M} \log p(\mathbf{y},\boldsymbol{\theta})\right\}}{\int \exp\left\{E_{-\boldsymbol{\theta}_M} \log p(\mathbf{y},\boldsymbol{\theta})\right\} d\boldsymbol{\theta}_M}$$

until the increase in $\underline{p}(\mathbf{y}; q)$ is negligible.

Connection with Gibbs Sampling

- An alternative expression for the q_i^* is

$$q_i^*(\boldsymbol{\theta}_i) \propto \exp\left\{E_{-\boldsymbol{\theta}_i} \log p(\boldsymbol{\theta}_i \mid \text{rest})\right\}$$

$$\text{rest} \equiv \{\mathbf{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_M\}$$

- The distributions $\boldsymbol{\theta}_i \mid \text{rest}$ are called full conditionals in Markov Chain Monte Carlo
- Gibbs sampling uses repeated draws from these
- In fact, product density transforms and Gibbs are tractable in the same scenarios

Product DT Example 1: Normal Random Sample

- Objective is to approximate Bayesian inference for a random sample from a Normal distribution

$$X_i \mid \mu, \sigma^2 \overset{\text{ind.}}{\sim} N\left(\mu, \sigma^2\right)$$

with conjugate priors

$$\mu \sim N\left(\mu_\mu, \sigma_\mu^2\right) \quad \text{and} \quad \sigma^2 \sim \text{IG}(A, B)$$

- The product density transform approximation of $p(\mu, \sigma^2 \mid \boldsymbol{x})$ is

$$q\left(\mu, \sigma^2\right) = q_\mu(\mu) q_{\sigma^2}\left(\sigma^2\right)$$

- The optimal densities take the form

$$q_\mu^*(\mu) \propto \exp\left[E_{\sigma^2}\left\{\log p\left(\mu \mid \sigma^2, \mathbf{x}\right)\right\}\right] \quad \text{and}$$

$$q_{\sigma^2}^*\left(\sigma^2\right) \propto \exp\left[E_\mu\left\{\log p\left(\sigma^2 \mid \mu, \mathbf{x}\right)\right\}\right]$$

Product DT Example 1: Normal Random Sample

- The resulting estimates are:

$$q^*_{\sigma^2}(\sigma^2) \text{ is InverseGamma} \left(A + \frac{n}{2}, B + \frac{1}{2} E_\mu \left\| \mathbf{x} - \mu \mathbf{1}_n \right\|^2 \right)$$

$$q^*_\mu(\mu) \text{ is Normal} \left(\frac{n\bar{X} E_{\sigma^2}\left(1/\sigma^2\right) + \mu_\mu/\sigma^2_\mu}{n E_{\sigma^2}\left(1/\sigma^2\right) + 1/\sigma^2_\mu}, \frac{1}{n E_{\sigma^2}\left(1/\sigma^2\right) + 1/\sigma^2_\mu} \right)$$

and

$$\begin{aligned} \log \underline{p}(\boldsymbol{x}; q) = & \frac{1}{2} - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log\left(\sigma^2_{q(\mu)}/\sigma^2_\mu\right) \\ & - \frac{\left(\mu_{q(\mu)} - \mu_\mu\right)^2 + \sigma^2_{q(\mu)}}{2\sigma^2_\mu} + A \log(B) \\ & - \left(A + \frac{n}{2}\right) \log\left(B_{q(\sigma^2)}\right) + \log \Gamma\left(A + \frac{n}{2}\right) - \log \Gamma(A) \end{aligned}$$

- This leads to the algorithm on the next slide

Product DT Example 1: Normal Random Sample

Algorithm 2 Iterative scheme for obtaining the parameters in the optimal densities q_{μ}^{*} and $q_{\sigma^2}^{*}$ in the Normal random sample example.

Initialize: $B_{q(\sigma^2)} > 0$

Cycle:

$$\sigma^2_{q(\mu)} \leftarrow \left\{ n \left(A + \tfrac{n}{2} \right) / B_{q(\sigma^2)} + 1/\sigma^2_{\mu} \right\}^{-1},$$

$$\mu_{q(\mu)} \leftarrow \left\{ n\bar{X} \left(A + \tfrac{n}{2} \right) / B_{q(\sigma^2)} + \mu_{\mu}/\sigma^2_{\mu} \right\} \sigma^2_{q(\mu)},$$

$$B_{q(\sigma^2)} \leftarrow B + \tfrac{1}{2} \left(\left\| \mathbf{x} - \mu_{q(\mu)} \mathbf{1}_n \right\|^2 + n\sigma^2_{q(\mu)} \right)$$

until the increase in $\underline{p}(\mathbf{x}; q)$ is negligible.

Product DT Example 1: Normal Random Sample

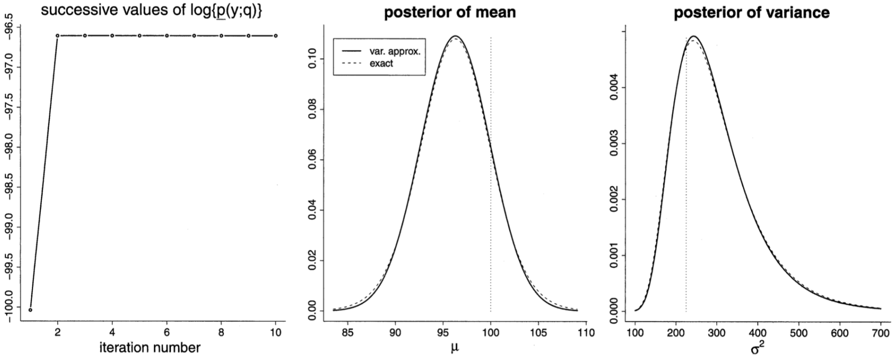
- Upon convergence, the posterior densities are approximated as

$$p(\mu \mid \boldsymbol{x}) \approx \left\{2\pi\left(\sigma^2_{q(\mu)}\right)^*\right\}^{-1/2} \exp\left[-\left(\mu - \mu^*_{q(\mu)}\right)^2 / \left\{2\left(\sigma^2_{q(\mu)}\right)^*\right\}\right]$$

$$p\left(\sigma^2 \mid \boldsymbol{x}\right) \approx \frac{\left(B^*_{q(\sigma^2)}\right)^{A+\frac{n}{2}}}{\Gamma\left(A+\frac{n}{2}\right)} \left(\sigma^2\right)^{-A-\frac{n}{2}-1} \exp\left(B^*_{q(\sigma^2)}/\sigma^2\right), \quad \sigma^2 > 0$$

- Next slide's plots compare product density variational approximations with exact posterior density
  - Sample size $n = 20$ from $N(100, 225)$
  - Vague priors chosen: $\mu \sim N(0, 10^8)$, $\sigma^2 \sim IG(\frac{1}{100}, \frac{1}{100})$
  - Initial value $B_{q(\sigma^2)} = 1$
  - Convergence very rapid, accuracy quite good

Product DT Example 1: Normal Random Sample



Product DT Example 2: Linear Mixed Model

- Objective is to approximate Bayesian inference for a random sample from a Gaussian linear mixed model

$$\boldsymbol{y}|\beta, \boldsymbol{u}, \boldsymbol{G}, \boldsymbol{R} \sim N(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{R}), \quad \boldsymbol{u}|\boldsymbol{G} \sim N(\boldsymbol{0}, \boldsymbol{G})$$

  where
  - $\boldsymbol{y}$ is $n \times 1$ response,
  - β is $p \times 1$ fixed effects,
  - $\boldsymbol{u}$ is random effects,
  - $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices, and
  - $\boldsymbol{G}$ and $\boldsymbol{R}$ are covariance matrices
- Conjugate priors are

$$\beta \sim N\left(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}\right), \quad \sigma_{u\ell}^2 \sim \mathrm{IG}\left(A_{u\ell}, B_{u\ell}\right), 1 \leq \ell \leq r, \quad \sigma_\varepsilon^2 \sim \mathrm{IG}\left(A_\varepsilon, B_\varepsilon\right)$$

Product DT Example 2: Linear Mixed Model

- The two-component product transform is

$$q\left(\beta, \boldsymbol{u}, \sigma^2_{u1}, \ldots, \sigma^2_{ur}, \sigma^2_{\varepsilon}\right) = q_{\beta,\boldsymbol{u}}(\beta, \boldsymbol{u}) q_{\boldsymbol{\sigma}^2}\left(\sigma^2_{u1}, \ldots, \sigma^2_{ur}, \sigma^2_{\varepsilon}\right)$$

- This leads to optimal densities

$$q^*_{\beta,\boldsymbol{u}}(\beta, \boldsymbol{u}) \text{ is a Multivariate Normal density}$$

$$q^*_{\boldsymbol{\sigma}^2} \text{ is a product of } r+1 \text{ Inverse Gamma densities}$$

and. . .

Product DT Example 2: Linear Mixed Model

$$
\begin{aligned}
&\log \underline{p}(\mathbf{y}; q) \\
&= \tfrac{1}{2}\left(p + \textstyle\sum_{\ell=1}^{r} K_\ell\right) - \tfrac{n}{2}\log(2\pi) - \tfrac{p}{2}\log\left(\sigma_\beta^2\right) \\
&\quad + \tfrac{1}{2}\log\left|\boldsymbol{\Sigma}_{q(\beta,\mathbf{u})}\right| - \tfrac{1}{2\sigma_\beta^2}\left\{\left\|\boldsymbol{\mu}_{q(\beta)}\right\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\beta)}\right)\right\} \\
&\quad + A_\varepsilon \log\left(B_\varepsilon\right) - \left(A_\varepsilon + \tfrac{n}{2}\right)\log\left(B_{q(\sigma_\varepsilon^2)}\right) \\
&\quad + \log\Gamma\left(A_\varepsilon + \tfrac{n}{2}\right) - \log\Gamma\left(A_\varepsilon\right) \\
&\quad + \textstyle\sum_{\ell=1}^{r}\left\{A_{u\ell}\log\left(B_{u\ell}\right) - \left(A_{u\ell} + \tfrac{K_\ell}{2}\right)\log\left(B_{q\left(\sigma_{u\ell}^2\right)}\right)\right. \\
&\quad \left. + \log\Gamma\left(A_{u\ell} + \tfrac{K_\ell}{2}\right) - \log\Gamma\left(A_{u\ell}\right)\right\}
\end{aligned}
$$

- This leads to the algorithm on the next slide

Algorithm 3 Iterative scheme for obtaining the parameters in the optimal densities $q^*_{\beta,\boldsymbol{u}}$ and $q^*_{\boldsymbol{\sigma}^2}$ in the Bayesian linear mixed model example.

Product DT Example 2: Linear Mixed Model

Initialize: $B_{q(\sigma_\varepsilon^2)}, B_{q\left(\sigma_{u1}^2\right)}, \ldots, B_{q(\sigma_{ur}^2)} > 0$

Cycle:

$$\boldsymbol{\Sigma}_{q(\beta,\mathbf{u})} \leftarrow \left\{ \frac{A_\varepsilon + \frac{n}{2}}{B_{q(\sigma_\varepsilon^2)}} \mathbf{C}^T\mathbf{C} + \text{ blockdiag } \left(\sigma_\beta^{-2}\mathbf{I}_p, \frac{A_{u1} + \frac{1}{2}K_1}{B_{q\left(\sigma_{u1}^2\right)}}\mathbf{I}_{K_1}, \ldots, \frac{A_{ur} + \frac{1}{2}K_r}{B_{q(\sigma_{ur}^2)}}\mathbf{I}_{K_r} \right) \right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\beta,\mathbf{u})} \leftarrow \left(\frac{A_\varepsilon + \frac{n}{2}}{B_{q\left(\sigma_\varepsilon^2\right)}} \right) \boldsymbol{\Sigma}_{q(\beta,\mathbf{u})}\mathbf{C}^T\mathbf{y}$$

$$B_{q(\sigma_\varepsilon^2)} \leftarrow B_\varepsilon + \frac{1}{2}\left\{ \left\| \mathbf{y} - \mathbf{C}\boldsymbol{\mu}_{q(\beta,\mathbf{u})} \right\|^2 + \operatorname{tr}\left(\mathbf{C}^T\mathbf{C}\boldsymbol{\Sigma}_{q(\beta,\mathbf{u})} \right) \right\}$$

$$B_{q\left(\sigma_{u\ell}^2\right)} \leftarrow B_{u\ell} + \frac{1}{2}\left\{ \left\| \boldsymbol{\mu}_{q(\mathbf{u}_\ell)} \right\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_q\left(\mathbf{u}_\ell\right)\right) \right\} \quad \text{for } 1 \le \ell \le r$$

until the increase in $\underline{p}(\mathbf{x}; q)$ is negligible.

Product DT Example 2: Linear Mixed Model

- Upon convergence, the posterior densities are approximated as

$$p(\beta, \mathbf{u} \mid \mathbf{y}) \approx \text{ the } N\left(\boldsymbol{\mu}^*_{q(\beta,\mathbf{u})}, \boldsymbol{\Sigma}^*_{q(\beta,\mathbf{u})}\right) \text{ density function}$$

$$p\left(\sigma^2_{u1}, \ldots, \sigma^2_{ur}, \sigma^2_{\varepsilon} \mid \boldsymbol{y}\right) \approx$$

$$\text{product of the IG}\left(A_{u\ell} + \tfrac{1}{2}K_{\ell}, B^*_{q\left(\sigma^2_{u\ell}\right)}\right), 1 \leq \ell \leq r, \text{ density}$$

$$\text{functions together with the IG}\left(A_{\varepsilon} + \tfrac{1}{2}n, B^*_{q(\sigma^2_{\varepsilon})}\right) \text{ density function}$$

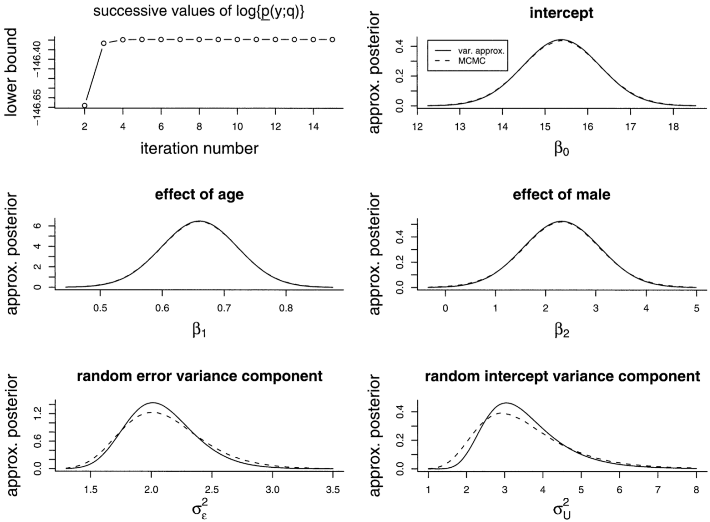
Product DT Example 2: Linear Mixed Model

- Next slide's plots compare product density variational approximations with exact posterior density
  - Data set is longitudinal orthodontic measurements (Pinheiro & Bates, 2000)
  - Random intercept model:

$$
\begin{aligned}
&\text{distance}_{ij} \mid U_i \overset{\text{ind.}}{\sim} N\left(\beta_0 + U_i + \beta_1 \text{ age}_{ij} + \beta_2 \text{male}_i, \sigma_\varepsilon^2\right) \\
&U_i \mid \sigma_u^2 \overset{\text{ind.}}{\sim} N\left(0, \sigma_u^2\right), \quad 1 \le i \le 27, 1 \le j \le 4, \\
&\beta_i \overset{\text{ind.}}{\sim} N\left(0, \sigma_\beta^2\right), \quad \sigma_u^2, \sigma_\varepsilon^2 \overset{\text{ind}}{\sim} \mathrm{IG}(A, B)
\end{aligned}
$$

$+$ Vague priors chosen: $\sigma_\beta^2 = 10^8,\ A = B = \frac{1}{100}$ $+$ Compared against kernel density estimates using 1M MCMC samples $+$ Convergence again quite rapid, estimates quite close to MCMC, statistical significance of all parameters

Product DT Example 2: Linear Mixed Model



Parametric DT Example: Poisson Regression

- Now assume q is subject to the parametric restriction
  - Belongs to a specific parametric family that (hopefully) results in a more tractable approximation to the posterior density
- Poisson Regression with Gaussian Transform example:
  - Consider the Bayesian Poisson Regression Model:

$$Y_i|\beta_0...\beta_k \sim Poisson(exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_1 x_{ki}))$$

  - With priors on the coefficient vector of $\beta \sim N(\mu_\beta, \Sigma_\beta)$

Parametric DT Example: Poisson Regression

- The likelihood is

$$p(\boldsymbol{y}|\beta) = exp\{\boldsymbol{y}^T \boldsymbol{X}\beta - \boldsymbol{I}_n^T exp(\boldsymbol{X}\beta) - \boldsymbol{I}_n^T log(\boldsymbol{y}!)\}$$

- This leads to an integral that has no closed form solution (intractable):

$$\begin{aligned}
&p(\boldsymbol{y}) = (2\pi)^{-(k+1)/2}|\boldsymbol{\Sigma}_\beta|^{-1/2} \\
&\int_{\mathbb{R}^{k+1}} exp\{\boldsymbol{y}^T \boldsymbol{X}\beta - \boldsymbol{I}_n^T exp(\boldsymbol{X}\beta) - \boldsymbol{I}_n^T log(\boldsymbol{y}!) \\
&-\frac{1}{2}(\beta - \boldsymbol{\mu}_\beta)^T \boldsymbol{\Sigma}_\beta{}^{-1}(\beta - \boldsymbol{\mu}_\beta)\} d\beta
\end{aligned}$$

Parametric DT Example: Poisson Regression

- Take $q \sim N(\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)})$

$$q(\beta; \boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}_{q(\beta)}|^{-1/2} * exp\{-\tfrac{1}{2}(\beta - \boldsymbol{\mu}_{q(\beta)})^T \boldsymbol{\Sigma}_{q(\beta)}^{-1}(\beta -$$

- Then the lower bound as defined earlier gives explicitly

$$\begin{aligned}
&\log \underline{p}\left(\mathbf{y}; \boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)}\right) \\
&= \mathbf{y}^T \mathbf{X} \boldsymbol{\mu}_{q(\beta)} - \mathbf{1}_n^T \exp\left\{\mathbf{X}\boldsymbol{\mu}_{q(\beta)} + \frac{1}{2} \text{diagonal}\left(\mathbf{X}\boldsymbol{\Sigma}_{q(\beta)}\mathbf{X}^T\right)\right\} \\
&\quad - \frac{1}{2}\left(\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta\right)^T \boldsymbol{\Sigma}_\beta^{-1}\left(\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta\right) - \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\Sigma}_{q(\beta)}\right) \\
&\quad + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{q(\beta)}\right| - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_\beta\right| + \frac{k+1}{2} \\
&\quad - \mathbf{1}_n^T \log(\mathbf{y}!)
\end{aligned}$$

Parametric DT Example: Poisson Regression

- From earlier,

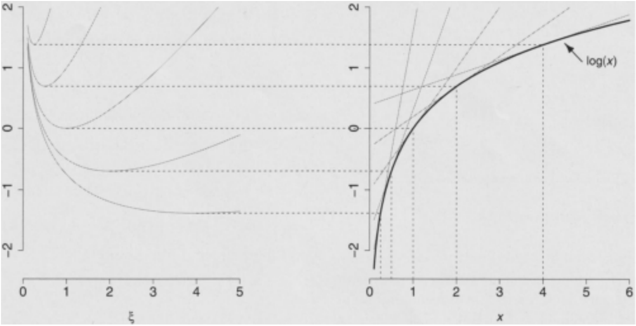
$$\log p(y) \geq \log \underline{p}(\boldsymbol{y}; \boldsymbol{\mu}_{\boldsymbol{q(\beta)}}, \boldsymbol{\Sigma}_{\boldsymbol{q(\beta)}})$$

- The optimal variational parameters are found through maximizing this inequality using Newton-Raphson iteration
- This minimizes the K-L divergence and provides the optimal Gaussian density transform q^* as $N(\boldsymbol{\mu}^*_{q(\beta)}, \boldsymbol{\Sigma}^*_{q(\beta)})$

Tangent Transform Approach

- Not all variational approximations fit into Kullback-Leibler divergence framework
- Tangent transforms work with *tangent-type* representations of concave/convex functions (Underpinned by Theory of Convex Duality)

$$log(x) = min\{\xi x - log(\xi) - 1\} \text{ for all } x > 0$$



TT Example: Bayesian Logistic Regression

- Consider Bayesian logistic regression model

  $Y_i|\beta_0, ..., \beta_k \sim Bernoulli([1+exp\{-(\beta_0+\beta_1 x_{1i}+...+\beta_k x_{ki}\}]^{-1}),\ 1 \leq i \leq$

- With priors on the coefficient vector of $\beta \sim N(\mu_\beta, \Sigma_\beta)$
- The likelihood is

$$p(\boldsymbol{y}|\beta) = exp[\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X}\beta - \boldsymbol{I}_{\boldsymbol{n}}^{\boldsymbol{T}} log\{I_n + exp(\boldsymbol{X}\beta)\}]$$

and the posterior density of β is

$$p(\beta|\boldsymbol{y}) = p(\boldsymbol{y}, \beta) / \int_{\mathbb{R}^{k+1}} p(\boldsymbol{y}, \beta) \boldsymbol{d}\beta$$

where

$$p(y|\beta) = exp[\boldsymbol{y}^T \boldsymbol{X}\beta - \boldsymbol{I}_{\boldsymbol{n}}^{\boldsymbol{T}} log(\boldsymbol{I}_{\boldsymbol{n}} + exp(\boldsymbol{X}\beta)) \\ -\tfrac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)^T \boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta) - \tfrac{k+1}{2} log(2\pi) - \tfrac{1}{2} log|\boldsymbol{\Sigma}_\beta|]$$

and the denominator is an intractable integral

TT Example: Bayesian Logistic Regression

- It can be shown $-\log(1+e^x)$ are the maxima of a family of parabolas:

$$-\log(1+e^x) = max\{A(\xi)x^2 - \frac{1}{2}x + C(\xi)\}$$

$$A(\xi) = -tanh(\xi/2)/(4\xi)$$

$$C(\xi) = \xi/2 - log(1+e^\xi) + \xi tanh(\xi/2)/4$$

- above equation is a *tangent-type* representation of a convex function.

TT Example: Bayesian Logistic Regression

- Under similar derivations as the previous problem, the family of variational approximations to $\beta|y$ is:

$$\beta|\boldsymbol{y};\boldsymbol{\xi} \sim N(\mu(\xi), \Sigma(\xi))$$

where

$$\boldsymbol{\Sigma}(\boldsymbol{\xi}) = [\boldsymbol{\Sigma}_{\beta}^{-1} - 2\boldsymbol{X}^{\boldsymbol{T}} diag A(\xi)\boldsymbol{X}]^{-1}$$

and

$$\boldsymbol{\mu}(\boldsymbol{\xi}) = \boldsymbol{\Sigma}(\xi)\{\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{y} - \frac{1}{2}\boldsymbol{I}) + \boldsymbol{\Sigma}_{\beta}^{-1}\boldsymbol{\mu}_{\beta}\}$$

- Note: Jaakkola and Jordan derived a simpler and quicker algorithm for maximizing numerically based on Expectation Maximization

Conclusion

- Doesn't cover the accuracy of the approximations
  - makes it hard to compare with MCMC
- Usefulness of the variational approximations increases as the size increases
  - MCMC begins to become untenable