MAST90125: Bayesian Statistical learning

Lecture 20: Variational Bayes

Feng Liu and Guoqi Qian

THE UNIVERSITY OF MELBOURNE

Computational techniques discussed so far

- We have learned a variety of computing techniques that could be used to approximate the posterior distribution. As a reminder, these are
  - Direct approximation (requires you to define a grid of points)
  - Rejection sampling (requiring an envelope density)
  - Importance sampling (requiring an candidate/envelope density)

  and some more popular MCMC methods,
  - Metropolis-Hastings (requiring a proposal conditional/transition distribution)
  - Gibbs sampling (requiring conditional posteriors).
- A consistent feature of these methods is heavy computing demands. In this lecture, we will introduce some approximate methods that aim to minimise computational cost.

Theory of Variational Bayes

- Consider a posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y})$, where $\boldsymbol{\theta}$ is the set of parameters and $\mathbf{y}$ is data. Variational Bayesian inference aims to approximate $p(\boldsymbol{\theta}|\mathbf{y})$ with a simpler probability distribution $Q(\boldsymbol{\theta})$. In particular, we will focus on mean-field Variational Bayes, where $Q(\boldsymbol{\theta})$ can be expressed as the factorisation $\prod_{j=1}^{K} Q(\boldsymbol{\theta}_j)$, where K is the number of disjoint sub-vectors we have partitioned the parameter vectors $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ into.

- To determine the best choice for $Q(\boldsymbol{\theta})$, we need to minimise the Kullback-Leibler divergence between the posterior $p(\boldsymbol{\theta} \mid \mathbf{y})$ and $Q(\boldsymbol{\theta})$,

$$D_{KL}\{Q(\boldsymbol{\theta})||p(\boldsymbol{\theta} \mid \mathbf{y})\} = \int_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})\{\log(Q(\boldsymbol{\theta})) - \log(p(\boldsymbol{\theta} \mid \mathbf{y}))\}\mathrm{d}\boldsymbol{\theta},$$

which we previously encountered when justifying various model selection methods.

Theory of Variational Bayes

- To determine the best choice for $Q(\boldsymbol{\theta})$, we need to minimise the Kullback-Leibler divergence between the posterior $p(\boldsymbol{\theta} \mid \mathbf{y})$ and $Q(\boldsymbol{\theta})$,

$$D_{KL}\{Q(\boldsymbol{\theta})||p(\boldsymbol{\theta} \mid \mathbf{y})\} = \int_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})\{\log(Q(\boldsymbol{\theta})) - \log(p(\boldsymbol{\theta} \mid \mathbf{y}))\}\mathrm{d}\boldsymbol{\theta},$$

  which we previously encountered when justifying various model selection methods.
- Substituting the factorised version of the approximate distribution $\prod_{j=1}^{K} Q(\boldsymbol{\theta}_j)$ and applying Bayes rule to $p(\boldsymbol{\theta}|\mathbf{y})$, $D_{KL}\{Q(\boldsymbol{\theta})||p(\boldsymbol{\theta} \mid \mathbf{y})\}$ can be written as,

$$D_{KL}\{Q(\boldsymbol{\theta})||p(\boldsymbol{\theta} \mid \mathbf{y})\} = \int_{\boldsymbol{\theta}} \Big\{\prod_{j=1}^{K} Q(\boldsymbol{\theta}_j)\Big\}\Big\{\sum_{j=1}^{K} \log(Q(\boldsymbol{\theta}_j)) - \log(p(\boldsymbol{\theta}, \mathbf{y}))\Big\}\mathrm{d}\boldsymbol{\theta} + \log(p(\mathbf{y}))$$

Theory of Variational Bayes

- Substituting the factorised version of the approximate distribution $\prod_{j=1}^{K} Q(\boldsymbol{\theta}_j)$ and applying Bayes rule to $p(\boldsymbol{\theta}|\mathbf{y})$, $D_{KL}\{Q(\boldsymbol{\theta})||p(\boldsymbol{\theta} \mid \mathbf{y})\}$ can be written as,

$$D_{KL}\{Q(\boldsymbol{\theta})||p(\boldsymbol{\theta} \mid \mathbf{y})\} = \int_{\boldsymbol{\theta}} \left\{ \prod_{j=1}^{K} Q(\boldsymbol{\theta}_j) \right\} \left\{ \sum_{j=1}^{K} \log(Q(\boldsymbol{\theta}_j)) - \log(p(\boldsymbol{\theta}, \mathbf{y})) \right\} \mathrm{d}\boldsymbol{\theta} + \log(p(\mathbf{y}))$$

- Because $\log(p(\mathbf{y}))$ is a constant with respect to $\boldsymbol{\theta}$, the term requiring minimisation reduces to,

$$L(\boldsymbol{\theta}) = \int_{\boldsymbol{\theta}} \left\{ \prod_{i=1}^{K} Q(\boldsymbol{\theta}_i) \right\} \left\{ \sum_{j=1}^{K} \log(Q(\boldsymbol{\theta}_j)) - \log(p(\boldsymbol{\theta}, \mathbf{y})) \right\} \mathrm{d}\boldsymbol{\theta}, \qquad (1)$$

noting that changing the index on the product from j to i does not result in loss of generality.

Theory of Variational Bayes

- We now can minimise $D_{KL}\{Q(\boldsymbol{\theta})||p(\boldsymbol{\theta} \mid \mathbf{y})\}$ with respect to each $\boldsymbol{\theta}_j; j = 1, \ldots, K$, rather than $\boldsymbol{\theta}$. To determine the function to minimise $L(\boldsymbol{\theta}_j)$, we need to simplify $L(\boldsymbol{\theta}_j)$ from (1).

- We analyse (1) by the following two parts:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \int_{\boldsymbol{\theta}} \Big\{ \prod_{i=1}^{K} Q(\boldsymbol{\theta}_i) \Big\} \Big\{ \sum_{j=1}^{K} \log(Q(\boldsymbol{\theta}_j)) - \log(p(\boldsymbol{\theta}, \mathbf{y})) \Big\} \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \Big\{ \prod_{i=1}^{K} Q(\boldsymbol{\theta}_i) \Big\} \sum_{j=1}^{K} \log(Q(\boldsymbol{\theta}_j)) \mathrm{d}\boldsymbol{\theta} \quad \rightarrow (\text{part I}) \\ &\quad - \int_{\boldsymbol{\theta}} \Big\{ \prod_{i=1}^{K} Q(\boldsymbol{\theta}_i) \Big\} \log(p(\boldsymbol{\theta}, \mathbf{y})) \mathrm{d}\boldsymbol{\theta} \quad \rightarrow (\text{part II}) \end{aligned} \tag{2}$$

Theory of Variational Bayes

- For $\int_{\boldsymbol{\theta}} \left\{ \prod_{i=1}^{K} Q(\boldsymbol{\theta}_i) \right\} \sum_{j=1}^{K} \log(Q(\boldsymbol{\theta}_j)) \mathrm{d}\boldsymbol{\theta}$, the simplification for $\boldsymbol{\theta}_j$ will drop the sum with respect to j and then separate out the part of the product indexed by j,

$$\begin{aligned}
\int_{\boldsymbol{\theta}} \left\{ \prod_{i=1}^{K} Q(\boldsymbol{\theta}_i) \right\} \log(Q(\boldsymbol{\theta}_j)) \mathrm{d}\boldsymbol{\theta} &= \int_{\boldsymbol{\theta}_j} \int_{\boldsymbol{\theta}_{-j}} Q(\boldsymbol{\theta}_j) \left\{ \prod_{i \neq j} Q(\boldsymbol{\theta}_i) \right\} \log(Q(\boldsymbol{\theta}_j)) \mathrm{d}\boldsymbol{\theta}_j \mathrm{d}\boldsymbol{\theta}_{-j} \\
&= \int_{\boldsymbol{\theta}_j} Q(\boldsymbol{\theta}_j) \log(Q(\boldsymbol{\theta}_j)) \mathrm{d}\boldsymbol{\theta}_j \left\{ \prod_{i \neq j} \int_{\boldsymbol{\theta}_i} Q(\boldsymbol{\theta}_i) \mathrm{d}\boldsymbol{\theta}_i \right\} \\
&= \int_{\boldsymbol{\theta}_j} Q(\boldsymbol{\theta}_j) \log(Q(\boldsymbol{\theta}_j)) \mathrm{d}\boldsymbol{\theta}_j. \qquad (3)
\end{aligned}$$

Theory of Variational Bayes

- To simplify the second part of (2), $\int_{\boldsymbol{\theta}} \left\{ \prod_{i=1}^{K} Q(\boldsymbol{\theta}_i) \right\} \log(p(\boldsymbol{\theta}, \mathbf{y}))\mathrm{d}\boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}_j$, we separate out the part of the product indexed by j,

$$\int_{\boldsymbol{\theta}_1} Q(\boldsymbol{\theta}_1) \times \ldots \times \int_{\boldsymbol{\theta}_K} Q(\boldsymbol{\theta}_K) \log(p(\boldsymbol{\theta}, \mathbf{y}))\mathrm{d}\boldsymbol{\theta}_1 \ldots \mathrm{d}\boldsymbol{\theta}_K = \int_{\boldsymbol{\theta}_j} Q(\boldsymbol{\theta}_j) E_{-j}\{\log(p(\boldsymbol{\theta}, \mathbf{y}))\}\mathrm{d}\boldsymbol{\theta}_j. \quad (4)$$

From the results in (3) and (4), we can determine that $L(\boldsymbol{\theta})$ is a function of $L(\boldsymbol{\theta}_j)$ and a constant (if $Q(\boldsymbol{\theta}_{-j})$ is fixed),

$$\begin{aligned} L(\boldsymbol{\theta}) &= \int_{\boldsymbol{\theta}_j} Q(\boldsymbol{\theta}_j) \Big\{ \log(Q(\boldsymbol{\theta}_j)) - E_{-j}\{\log(p(\boldsymbol{\theta}, \mathbf{y}))\} \Big\} \mathrm{d}\boldsymbol{\theta}_j + \sum_{i \neq j} \int_{\boldsymbol{\theta}_i} Q(\boldsymbol{\theta}_i) \log(Q(\boldsymbol{\theta}_i))\mathrm{d}\boldsymbol{\theta}_i \\ &= L(\boldsymbol{\theta}_j) + \text{Constant}. \end{aligned}$$

Theory of Variational Bayes

- We can re-write $L(\boldsymbol{\theta}_j)$ as follows,

$$\begin{aligned} L(\boldsymbol{\theta}_j) &= \int_{\boldsymbol{\theta}_j} Q(\boldsymbol{\theta}_j)\Big\{ \log(Q(\boldsymbol{\theta}_j)) - E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}\Big\}\mathrm{d}\boldsymbol{\theta}_j \\ &= \int_{\boldsymbol{\theta}_j} Q(\boldsymbol{\theta}_j)\Big\{ \log(Q(\boldsymbol{\theta}_j)) - \log(e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}})\Big\}\mathrm{d}\boldsymbol{\theta}_j. \end{aligned} \tag{5}$$

- $L(\boldsymbol{\theta}_j)$ in (5) would correspond to the Kullback-Leibler divergence, except $e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}$ is not a valid probability distribution. If $e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}$ is finite integrable, we can define $Q^*(\boldsymbol{\theta}_j)$,

$$Q^*(\boldsymbol{\theta}_j) = \frac{e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}}{\int_{\boldsymbol{\theta}_j} e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}\mathrm{d}\boldsymbol{\theta}_j}, \quad \text{such that} \int_{\boldsymbol{\theta}_j} Q^*(\boldsymbol{\theta}_j)\mathrm{d}\boldsymbol{\theta}_j = 1. \tag{6}$$

Theory of Variational Bayes

- Let $z = \int_{\boldsymbol{\theta}_j} e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}} \mathrm{d}\boldsymbol{\theta}_j$. We can write $L(\boldsymbol{\theta}_j)$ to be proportional to the Kullback-Leibler divergence between $Q(\boldsymbol{\theta}_j)$ and $Q^*(\boldsymbol{\theta}_j)$ as defined in (6),

$$\begin{aligned} L(\boldsymbol{\theta}_j) &= \int_{\boldsymbol{\theta}_j} Q(\boldsymbol{\theta}_j)\Big\{\log(Q(\boldsymbol{\theta}_j)) - \log(e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}) - \log(z) + \log(z)\Big\}\mathrm{d}\boldsymbol{\theta}_j \\ &= \int_{\boldsymbol{\theta}_j} Q(\boldsymbol{\theta}_j)\Big\{\log(Q(\boldsymbol{\theta}_j)) - \log(Q^*(\boldsymbol{\theta}_j)) - \log(z)\Big\}\mathrm{d}\boldsymbol{\theta}_j \\ &= D_{KL}\{Q(\boldsymbol{\theta}_j)||Q^*(\boldsymbol{\theta}_j)\} - \log(z), \end{aligned}$$

which by definition must be minimised when $Q(\boldsymbol{\theta}_j) = Q^*(\boldsymbol{\theta}_j)$.

Find the best Q($\boldsymbol{\theta}$)

What we have got so far?

- Given a j, if $Q(\boldsymbol{\theta}_{-j})$ is fixed, we know the best $Q(\boldsymbol{\theta}_j)$ will be $Q^*(\boldsymbol{\theta}_j)$:

$$Q^*(\boldsymbol{\theta}_j) = \frac{e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}}{\int_{\boldsymbol{\theta}_j} e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}} \mathrm{d}\boldsymbol{\theta}_j}. \tag{7}$$

Find the best Q($\boldsymbol{\theta}$)

What we have got so far?

- Given a j, if $Q(\boldsymbol{\theta}_{-j})$ is fixed, we know the best $Q(\boldsymbol{\theta}_j)$ will be $Q^*(\boldsymbol{\theta}_j)$:

$$Q^*(\boldsymbol{\theta}_j) = \frac{e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}}{\int_{\boldsymbol{\theta}_j} e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}} \mathrm{d}\boldsymbol{\theta}_j}. \tag{7}$$

- If the kernel of $e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}$ is recognisible, we can directly obtain the analytic form of $Q^*(\boldsymbol{\theta}_j)$.

However, how to obtain the $Q(\boldsymbol{\theta})$?

- We can find $Q(\boldsymbol{\theta})$ iteratively, like the Gibbs sampling or Expectation-Maximisation method.

Find the best Q($\boldsymbol{\theta}$)–Coordinate Ascent Variational Inference (CAVI)

Assuming $\boldsymbol{\theta}$ is a p-dimension vector, the total iteration number is T, and $\mathbf{y}$ represents all data we have.

- Initialize $\boldsymbol{\theta}$ to be $\boldsymbol{\theta}^{(0)}$;
- Determine the kernel function regarding $e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}$ (unnormalised $Q^*(\boldsymbol{\theta}_j)$ for each j);
- for $t = 1 : T$
    - for $j = 1 : p$
        - Obtain $Q^*(\boldsymbol{\theta}_j^{(t+1)})$ based on $\boldsymbol{\theta}_{-j}^{(t)}$;
        - Update $\boldsymbol{\theta}_j^{(t)}$: $\boldsymbol{\theta}_j^{(t)} = \boldsymbol{\theta}_j^{(t+1)}$;
- Return $Q^*(\boldsymbol{\theta}_j^{(T)})$.

Gain and Loss of Variational Bayes

- It may look like that all we have done is partition the parameter space and defined independent but approximate posteriors. Which you may think is not much different from Gibbs sampling.
- However how are the parameters of the approximate posteriors determined?
  - By taking the expectation of the log posterior with respect to all parameters except the parameter of interest.
  - We can obtain the analytic form of the approximate posteriors without massive sampling procedures.
- What is the loss?

Gain and Loss of Variational Bayes

- It may look like that all we have done is partition the parameter space and defined independent but approximate posteriors. Which you may think is not much different from Gibbs sampling.
- However how are the parameters of the approximate posteriors determined?
  - By taking the expectation of the log posterior with respect to all parameters except the parameter of interest.
  - We can obtain the analytic form of the approximate posteriors without massive sampling procedures.
- What is the loss?
  - The $Q(\boldsymbol{\theta})$ is not the true posterior.

Example of Variational Bayes: linear mixed model

- Previously, we determined conditional posteriors for a linear mixed model that could be used in a Gibbs sampler. We will now determine how to implement mean-field variational Bayes for this problem (this example is complex, **not** required for the exam).

- The specification of this model is,
  - $p(y|\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \mathbf{u}, \tau_e, \tau_u) = \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \frac{1}{\tau_e}\mathbf{I}_n)$.
  - $p(\boldsymbol{\beta}) = \prod_{j=1}^{p} p(\boldsymbol{\beta}_j) \propto 1$.
  - $p(\mathbf{u}) = \mathcal{N}(\mathbf{0}_q, \frac{1}{\tau_u}\mathbf{K})$.
  - $p(\tau_e) = \text{Ga}(\alpha_e, \gamma_e)$.
  - $p(\tau_u) = \text{Ga}(\alpha_u, \gamma_u)$.

  meaning that the joint distribution, $p(y, \boldsymbol{\beta}, \mathbf{u}, \tau_e, \tau_u|\mathbf{X}, \mathbf{Z})$ is

$$\left(\frac{\tau_e}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{\tau_e(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})}{2}} \times 1 \times \left(\frac{\tau_u}{2\pi}\right)^{\frac{q}{2}} \det(\mathbf{K})^{-1/2} e^{-\frac{\tau_u \mathbf{u}'\mathbf{K}^{-1}\mathbf{u}}{2}} \times \frac{\gamma_u^{\alpha_u} \tau_u^{\alpha_u - 1} e^{-\gamma_u \tau_u}}{\Gamma(\alpha_u)} \times \frac{\gamma_e^{\alpha_e} \tau_e^{\alpha_e - 1} e^{-\gamma_e \tau_e}}{\Gamma(\alpha_e)}.$$

Example of Variational Bayes: linear mixed model

- For the approximate posteriors, we will look at the partition $\tau_e, \tau_u, \boldsymbol{\beta}, \mathbf{u}$.
- As the approximate posterior is defined as, $Q^*(\boldsymbol{\theta}_j) = \frac{e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}}{\int_{\boldsymbol{\theta}_j} e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}\mathrm{d}\boldsymbol{\theta}_j}$, just like when we determined posteriors, we only need to identify the kernel, and then take the expectation of the log-kernel.
- In the following slide, we will show the part regarding $\boldsymbol{\beta}$. In the end of this Lecture, details regarding $\tau_e, \tau_u, \mathbf{u}$ are demonstrated (these details are **not** required for exam).

Example of Variational Bayes: linear mixed model

- For β, the kernel and log-kernel are respectively

Kernel:$e^{-\frac{\tau_e(\mathbf{y}-\mathbf{X}\beta-\mathbf{Zu})'(\mathbf{y}-\mathbf{X}\beta-\mathbf{Zu})}{2}}$ Log-kernel: $-\dfrac{\tau_e(\mathbf{y}-\mathbf{X}\beta-\mathbf{Zu})'(\mathbf{y}-\mathbf{X}\beta-\mathbf{Zu})}{2}$

- The expected log-kernel $E_{-\beta}$(Log-kernel) is

$$
\begin{aligned}
&= -E_{\tau_e}(\tau_e)\left(\frac{\mathbf{y}'\mathbf{y}+\beta'\mathbf{X}`\mathbf{X}\beta+E_{\mathbf{u}}(\mathbf{uZ}'\mathbf{Zu})}{2}-\mathbf{y}'\mathbf{X}\beta-\mathbf{y}'\mathbf{Z}E_{\mathbf{u}}(\mathbf{u})+E_{\mathbf{u}}(\mathbf{u})'\mathbf{Z}'\mathbf{X}\beta\right) \\
&\propto -E_{\tau_e}(\tau_e)\left(\frac{\beta'\mathbf{X}'\mathbf{X}\beta}{2}-(\mathbf{y}-\mathbf{Z}E_{\mathbf{u}}(\mathbf{u}))'\mathbf{X}\beta\right) \\
&= -E_{\tau_e}(\tau_e)\left(\frac{\beta'\mathbf{X}'\mathbf{X}\beta-2(\mathbf{y}-\mathbf{Z}E_{\mathbf{u}}(\mathbf{u}))'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta}{2}\right)
\end{aligned}
$$

- which indicates that the **approximate posterior** for β is

$$
\mathcal{N}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y}-\mathbf{Z}E_{\mathbf{u}}(\mathbf{u})),\frac{1}{E_{\tau_e}(\tau_e)}(\mathbf{X}'\mathbf{X})^{-1}\right).
$$

Estimating parameters

- After similar procedures, we can determine the approximate posteriors $Q^*(\boldsymbol{\beta})$, $Q^*(\mathbf{u})$, $Q^*(\tau_u)$, $Q^*(\tau_e)$:
  - $Q^*(\tau_e)$ follows a Gamma distribution whose parameters are related to $E(\mathbf{u}), \mathrm{Var}(\mathbf{u}), E(\boldsymbol{\beta}), \mathrm{Var}(\boldsymbol{\beta}), \alpha_e, \gamma_e, \mathbf{y}, \mathbf{X}, \mathbf{Z}$;
  - $Q^*(\tau_u)$ follows a Gamma distribution whose parameters are related to $E(\mathbf{u}), \mathrm{Var}(\mathbf{u}), \mathbf{K}, \alpha_u, \gamma_u$;
  - $Q^*(\boldsymbol{\beta})$ follows a Normal distribution whose parameters are related to $\mathbf{y}, \mathbf{X}, \mathbf{Z}, E(\mathbf{u}), E(\tau_e)$;
  - $Q^*(\mathbf{u})$ follows a Normal distribution whose parameters are related to $\mathbf{y}, \mathbf{X}, \mathbf{Z}, \mathbf{K}, E(\boldsymbol{\beta})^{(t)}, E(\tau_e), E(\tau_u)$.

Estimating parameters

- To estimate these parameters, we use the CAVI algorithm.
    - Pick initial values $E(\boldsymbol{\beta})^{(0)}$, $\text{Var}(\boldsymbol{\beta})^{(0)}$, $E(\mathbf{u})^{(0)}$, $\text{Var}(\mathbf{u})^{(0)}$, $E(\tau_u)^{(0)}$, $E(\tau_e)^{(0)}$.
    - For $t = 1, 2, \ldots$
        - Calculate $E(\boldsymbol{\beta})^{(t)}$ from $\mathbf{y}, \mathbf{X}, \mathbf{Z}, E(\mathbf{u})^{(t-1)}$.
        - Calculate $\text{Var}(\boldsymbol{\beta})^{(t)}$ from $\mathbf{X}, E(\tau_e)^{(t-1)}$.
        - Calculate $E(\mathbf{u})^{(t)}, \text{Var}(\mathbf{u})^{(t)}$ from $\mathbf{y}, \mathbf{X}, \mathbf{Z}, \mathbf{K}, E(\boldsymbol{\beta})^{(t)}, E(\tau_e)^{(t-1)}, E(\tau_u)^{(t-1)}$.
        - Calculate $E(\tau_u)^{(t)}$ from $E(\mathbf{u})^{(t)}, \text{Var}(\mathbf{u})^{(t)}, \mathbf{K}, \alpha_u, \gamma_u$
        - Calculate $E(\tau_e)^{(t)}$ from $E(\mathbf{u})^{(t)}, \text{Var}(\mathbf{u})^{(t)}, E(\boldsymbol{\beta})^{(t)}, \text{Var}(\boldsymbol{\beta})^{(t)}, \alpha_e, \gamma_e, \mathbf{y}, \mathbf{X}, \mathbf{Z}$
    - Stop once convergence has been reached.
- Finally, we can obtain $Q^*(\boldsymbol{\beta})$, $Q^*(\mathbf{u})$, $Q^*(\tau_u)$, $Q^*(\tau_e)$.

Case study

- Consider the following dataset from animal breeding.

  A farm has two paddocks. The two paddocks carry 24 animals. The farmer wishes to increase the average weaning weight of their animals. To do so, the farmer wants to determine the genetic worth of the animals, in order to determine which animals should be mated together. The farmer also has pedigree records recording the parentage. To determine genetic worth, the following model is proposed:

  $$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon \quad \epsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma_e^2 \mathbf{I}_n), \mathbf{u} \sim \mathcal{N}(\mathbf{0}_q, \sigma_u^2 \mathbf{K}).$$

  where

  - $\mathbf{X}$ is an incidence matrix for paddock effects.
  - $\mathbf{Z}$ is an incidence matrix linking parents to children.
  - $\mathbf{K}$ is a kinship matrix determined for the parents.

- We coded Gibbs samplers for a linear mixed model in previous labs.
- In this example, we will compare the results obtained by fitting a Gibbs sampler to the Variational Bayes approximation.

Case study

- To conclude this lecture, we will now turn to R and run this example.
- The data can be downloaded from LMS as `farmdata.txt`. The relationship matrix $\mathbf{K}$ can also be downloaded from LMS as `Kmat.csv`
- Note: By comparing Variational Bayes to the Gibbs sampler, general comments about the performance of Variational Bayes will be made in the R code.

*(Details) Example of Variational Bayes: linear mixed model

- For the approximate posteriors, we will look at the partition $\tau_e, \tau_u, \boldsymbol{\beta}, \mathbf{u}$.

- As the approximate posterior is defined as, $Q^*(\boldsymbol{\theta}_j) = \frac{e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}}}{\int_{\boldsymbol{\theta}_j} e^{E_{-j}\{\log(p(\boldsymbol{\theta},\mathbf{y}))\}} \mathrm{d}\boldsymbol{\theta}_j}$, just like when we determined posteriors, we only need to identify the kernel, and then take the expectation of the log-kernel.

- For τ_e, the kernel and log-kernel are respectively

$$\begin{aligned}
\text{Kernel:} \qquad & \tau_e^{\frac{n}{2}} e^{-\frac{\tau_e(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Zu})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Zu})}{2}} \tau_e^{\alpha_e-1} e^{-\gamma_e\tau_e} \\
\text{Log-kernel:} \qquad & (n/2+\alpha_e-1)\log(\tau_e) - \tau_e(\gamma_e + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Zu})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Zu})/2)
\end{aligned}$$

- The expected log-kernel $E_{-\tau_e}(\text{Log-kernel})$ is

$$(\frac{n}{2}+\alpha_e-1)\log(\tau_e) - \tau_e(\gamma_e + \frac{\mathbf{y}'\mathbf{y} + E_{\boldsymbol{\beta}}(\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) + E_{\mathbf{u}}(\mathbf{u}'\mathbf{Z}'\mathbf{Z}\mathbf{u})}{2} - \mathbf{y}'\mathbf{X}E_{\boldsymbol{\beta}}(\boldsymbol{\beta}) - \mathbf{y}'\mathbf{Z}E_{\mathbf{u}}(\mathbf{u}) + E_{\mathbf{u}}(\mathbf{u})'\mathbf{Z}'\mathbf{X}E_{\boldsymbol{\beta}}(\boldsymbol{\beta}))$$

*(Details) Example of Variational Bayes: linear mixed model

- Using the tricks that $E(\mathbf{xx}') = \text{Var}(\mathbf{x}) + E(\mathbf{x})E(\mathbf{x})'$ and that $\mathbf{a}'\mathbf{Da} = \text{Tr}(\mathbf{a}'\mathbf{Da}) = \text{Tr}(\mathbf{Daa}')$, the expected log-kernel can be written as,

$$\left(\frac{n}{2} + \alpha_e - 1\right)\log(\tau_e) - \tau_e\left(\gamma_e + \frac{(\mathbf{y} - \mathbf{X}E_{\beta}(\boldsymbol{\beta}) - \mathbf{Z}E_{\mathbf{u}}(\mathbf{u}))'(\mathbf{y} - \mathbf{X}E_{\beta}(\boldsymbol{\beta}) - \mathbf{Z}E_{\mathbf{u}}(\mathbf{u})) + \text{Tr}(\mathbf{X}'\mathbf{X}\text{Var}(\boldsymbol{\beta})) + \text{Tr}(\mathbf{Z}'\mathbf{Z}\text{Var}(\mathbf{u}))}{2}\right)$$

- which indicates that the **approximate posterior** for τ_e is the distribution whose kernel is $\exp(E_{-\tau_e}(\text{Log-kernel}))$. That is,

$$\text{Ga}\left(\frac{n}{2} + \alpha_e, \gamma_e + \frac{(\mathbf{y} - \mathbf{X}E_{\beta}(\boldsymbol{\beta}) - \mathbf{Z}E_{\mathbf{u}}(\mathbf{u}))'(\mathbf{y} - \mathbf{X}E_{\beta}(\boldsymbol{\beta}) - \mathbf{Z}E_{\mathbf{u}}(\mathbf{u})) + \text{Tr}(\mathbf{X}'\mathbf{X}\text{Var}(\boldsymbol{\beta})) + \text{Tr}(\mathbf{Z}'\mathbf{Z}\text{Var}(\mathbf{u}))}{2}\right)$$

Note: To determine the approximate posterior is Gamma, if the kernel of a gamma distribution $x^{\alpha-1}e^{-\beta x}$ then the log-kernel is $(\alpha - 1)\log(x) - \beta x$. Similarly if the normal distribution kernel is $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ or equivalently $e^{-\frac{x^2-2\mu x}{2\sigma^2}}$, then the log-kernel is $-\frac{(x-\mu)^2}{2\sigma^2}$ or equivalently $-\frac{x^2-2\mu x}{2\sigma^2}$.

*(Details) Example of Variational Bayes: linear mixed model

- For τ_u, the kernel and log-kernel are respectively

$$\begin{aligned}
\text{Kernel:} \quad & \left(\frac{\tau_u}{2\pi}\right)^{\frac{q}{2}} e^{-\frac{\tau_u \mathbf{u}'\mathbf{K}^{-1}\mathbf{u}}{2}} \tau_u^{\alpha_u-1} e^{-\gamma_u \tau_u} \\
\text{Log-kernel:} \quad & (q/2+\alpha_u-1)\log(\tau_u) - \tau_u(\gamma_u + \mathbf{u}'\mathbf{K}^{-1}\mathbf{u}/2)
\end{aligned}$$

- The expected log-kernel $E_{-\tau_u}(\text{Log-kernel})$ is

$$\begin{aligned}
= \quad & (q/2+\alpha_u-1)\log(\tau_u) - \tau_u(\gamma_u + E_{\mathbf{u}}(\mathbf{u}'\mathbf{K}^{-1}\mathbf{u})/2) \\
= \quad & (q/2+\alpha_u-1)\log(\tau_u) - \tau_u(\gamma_u \mathrm{Tr}(\mathbf{K}^{-1}E_{\mathbf{u}}(\mathbf{u}\mathbf{u}'))/2) \\
= \quad & (q/2+\alpha_u-1)\log(\tau_u) - \tau_u(\gamma_u + E_{\mathbf{u}}(\mathbf{u})'\mathbf{K}^{-1}E_{\mathbf{u}}(\mathbf{u})/2 + \mathrm{Tr}(\mathbf{K}^{-1}Var(\mathbf{u}))/2)
\end{aligned}$$

- which indicates that the **approximate posterior** for τ_u is

$$\mathrm{Ga}\left(\frac{q}{2}+\alpha_u, \gamma_u + \frac{E_{\mathbf{u}}(\mathbf{u})'\mathbf{K}^{-1}E_{\mathbf{u}}(\mathbf{u}) + \mathrm{Tr}(\mathbf{K}^{-1}Var(\mathbf{u}))}{2}\right).$$

*(Details) Example of Variational Bayes: linear mixed model

- For $\mathbf{u}$, the kernel and log-kernel are respectively

$$\text{Kernel:} e^{-\frac{\tau_e(\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})'(\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})}{2}-\frac{\tau_u\mathbf{u}'\mathbf{K}^{-1}\mathbf{u}}{2}} \qquad \text{Log-kernel:} -\frac{\tau_e(\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})'(\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\mathbf{u})}{2}-\frac{\tau_u\mathbf{u}'\mathbf{K}^{-1}\mathbf{u}}{2}$$

- The expected log-kernel $E_{-\mathbf{u}}(\text{Log-kernel})$ is

$$\begin{aligned}
&= -E_{\tau_e}(\tau_e)\left(\frac{\mathbf{y}'\mathbf{y}+E_\beta(\beta'\mathbf{X}\mathbf{X}\beta)+\mathbf{u}\mathbf{Z}'\mathbf{Z}\mathbf{u}}{2}-\mathbf{y}'\mathbf{X}E_\beta(\beta)-\mathbf{y}'\mathbf{Z}\mathbf{u}+\mathbf{u}'\mathbf{Z}'\mathbf{X}E_\beta(\beta)\right)-\frac{E_{\tau_u}(\tau_u)\mathbf{u}'\mathbf{K}^{-1}\mathbf{u}}{2}\\
&\propto -\left(\frac{\mathbf{u}'(E_{\tau_e}(\tau_e)\mathbf{Z}'\mathbf{Z}+E_{\tau_u}(\tau_u)\mathbf{K}^{-1})\mathbf{u}}{2}-E_{\tau_e}(\tau_e)(\mathbf{y}-\mathbf{X}E_\beta(\beta))'\mathbf{Z}\mathbf{u}\right)\\
&= -\left(\frac{\mathbf{u}'(E_{\tau_e}(\tau_e)\mathbf{Z}'\mathbf{Z}+E_{\tau_u}(\tau_u)\mathbf{K}^{-1})\mathbf{u}-2E_{\tau_e}(\tau_e)(\mathbf{y}-\mathbf{X}E_\beta(\beta))'\mathbf{Z}(E_{\tau_e}(\tau_e)\mathbf{Z}'\mathbf{Z}+E_{\tau_u}(\tau_u)\mathbf{K}^{-1})^{-1}(E_{\tau_e}(\tau_e)\mathbf{Z}'\mathbf{Z}+E_{\tau_u}(\tau_u)\mathbf{K}^{-1})\mathbf{u}}{2}\right)
\end{aligned}$$

- which indicates that the **approximate posterior** for $\mathbf{u}$ is

$$\mathcal{N}\left(E_{\tau_e}(\tau_e)(E_{\tau_e}(\tau_e)\mathbf{Z}'\mathbf{Z}+E_{\tau_u}(\tau_u)\mathbf{K}^{-1})^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{X}E_\beta(\beta)),(E_{\tau_e}(\tau_e)\mathbf{Z}'\mathbf{Z}+E_{\tau_u}(\tau_u)\mathbf{K}^{-1})^{-1}\right).$$