Haziq Jamil
Department of Statistics
London School of Economics and Political Science

1 Variational inference for I-prior models

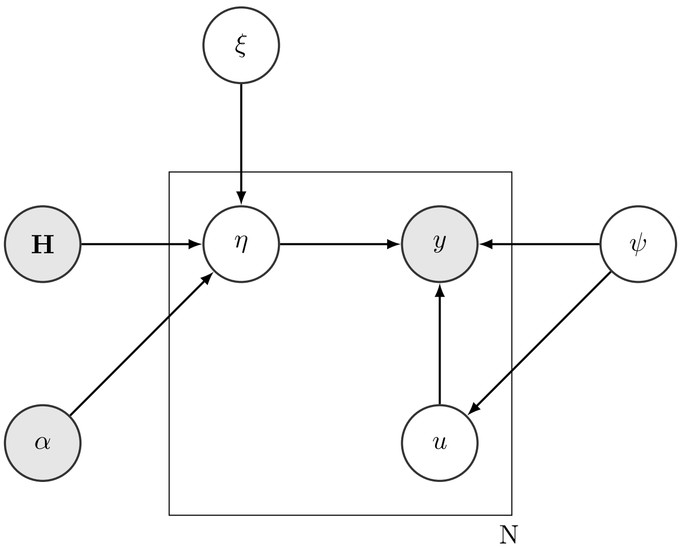
The model:

$$\mathbf{y} = \boldsymbol{\alpha} + \lambda \mathbf{H}\mathbf{w} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n)$$
$$\mathbf{w} \sim \mathrm{N}(\mathbf{0}, \psi\mathbf{I}_n)$$

Let $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n)$. Reparameterise $\xi = \lambda\psi$. Then $\psi^{-1}\mathbf{w}$ will have the same distribution as $\mathbf{u}$. Substituting this into the I-prior model, we have

$$\mathbf{y} = \boldsymbol{\alpha} + \xi \mathbf{H}\mathbf{u} + \boldsymbol{\epsilon}$$
$$\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n)$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n)$$

1.1 DAG for the I-prior model



1.2 Distributions

1.2.1 Priors

$$p(u_1, \dots, u_n) \equiv [\mathrm{N}(0, \psi^{-1})]^n$$
$$p(\xi, \psi) \propto \text{const.}$$

1.2.2 Joint data and latent

$$\begin{aligned} p(\mathbf{y}, \mathbf{u}, \xi, \psi) &= p(\mathbf{y}|\mathbf{u}, \xi, \psi) p(\mathbf{u}, \xi, \psi) \\ &= p(\mathbf{y}|\boldsymbol{\eta}) p(\mathbf{u}) p(\xi, \psi) \end{aligned}$$

where

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \xi \mathbf{H}\mathbf{u}$$

1.2.3 pdf/pmf

$$\begin{aligned} \log p(\mathbf{y}|\boldsymbol{\eta}) &= \log \mathrm{N}(\boldsymbol{\eta}, \psi^{-1}\mathbf{I}_n) \\ &= -\frac{n}{2}\log 2\pi + \frac{n}{2}\log\psi - \frac{\psi}{2}\|\mathbf{y} - \boldsymbol{\eta}\|^2 \\ &= -\frac{n}{2}\log 2\pi + \frac{n}{2}\log\psi - \frac{\psi}{2}\|\mathbf{y} - \boldsymbol{\alpha} - \xi\mathbf{H}\mathbf{u}\|^2 \\ &= -\frac{n}{2}\log 2\pi + \frac{n}{2}\log\psi - \frac{\psi}{2}\sum_{i=1}^n (y_i - \alpha - \xi\mathbf{H}_i\mathbf{u})^2 \end{aligned}$$

$$\begin{aligned} \log p(\mathbf{u}) &= \log \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n) \\ &= -\frac{n}{2}\log 2\pi + \frac{n}{2}\log\psi - \frac{\psi}{2}\|\mathbf{u}\|^2 \\ &= -\frac{n}{2}\log 2\pi + \frac{n}{2}\log\psi - \frac{\psi}{2}\sum_{i=1}^n u_i^2 \end{aligned}$$

1.3 Mean field approximation

$$q(\mathbf{u}, \xi, \psi) \equiv q(\mathbf{u})q(\xi)q(\psi)$$

1.3.1 Distribution of $\tilde{q}(\mathbf{u})$

$$
\begin{aligned}
\log \tilde{q}(\mathbf{u}) &= \mathrm{E}_{\xi,\psi}\left[-\frac{\psi}{2}\|\mathbf{y}-\boldsymbol{\alpha}-\xi\mathbf{H}\mathbf{u}\|^2 + \|\mathbf{u}\|^2\right] + \text{const.} \\
&= -\frac{\mathrm{E}\,\psi}{2}\,\mathrm{E}_{\xi,\psi}\left[\xi^2\mathbf{u}^\top\mathbf{H}^2\mathbf{u} + \mathbf{u}^\top\mathbf{u} - 2\xi(\mathbf{y}-\boldsymbol{\alpha})^\top\mathbf{H}\mathbf{u}\right] + \text{const.} \\
&= -\frac{\mathrm{E}\,\psi}{2}\Big(\mathbf{u}^\top(\mathrm{E}[\xi^2]\mathbf{H}^2 + \mathbf{I}_n)\mathbf{u} - 2\,\mathrm{E}\,\xi(\mathbf{y}-\boldsymbol{\alpha})^\top\mathbf{H}\mathbf{u}\Big) + \text{const.}
\end{aligned}
$$

Let $\mathbf{A} = \mathrm{E}[\xi^2]\mathbf{H}^2 + \mathbf{I}_n$ and $\mathbf{a} = \mathrm{E}[\xi]\mathbf{H}(\mathbf{y}-\boldsymbol{\alpha})$. Then, using the fact that

$$
\mathbf{u}^\top\mathbf{A}\mathbf{u} - 2\mathbf{a}^\top\mathbf{u} = (\mathbf{u}-\mathbf{A}^{-1}\mathbf{a})^\top\mathbf{A}(\mathbf{u}-\mathbf{A}^{-1}\mathbf{a}),
$$

we see the $\tilde{q}(\mathbf{u})$ is quadratic in $\mathbf{u}$, and we recognise this as the kernel of a multivariate normal density. Therefore,

$$
\tilde{q}(\mathbf{u}) \equiv \mathrm{N}(\mathbf{A}^{-1}\mathbf{a}, \mathbf{A}^{-1}/\,\mathrm{E}\,\psi)
$$

For convenience later in deriving the lower bound, we note that the second moment of $\tilde{q}(\mathbf{w})$ is equal to $\mathrm{E}[\mathbf{u}\mathbf{u}^\top] = \mathbf{A}^{-1}(\mathrm{E}^{-1}[\psi]\mathbf{I}_n + \mathbf{a}\mathbf{a}^\top\mathbf{A}^{-1}) =: \widetilde{\mathbf{U}}$.

1.3.2 Distribution of $\tilde{q}(\xi)$

$$
\begin{aligned}
\log \tilde{q}(\xi) &= \mathrm{E}_{\mathbf{u},\psi}\left[-\frac{\psi}{2}\|\mathbf{y}-\boldsymbol{\alpha}-\xi\mathbf{H}\mathbf{u}\|^2\right] + \text{const.} \\
&= -\frac{\mathrm{E}\,\psi}{2}\,\mathrm{E}_{\mathbf{u},\psi}\left[\xi^2\,\mathrm{tr}(\mathbf{H}^2\mathbf{u}\mathbf{u}^\top) - 2\xi(\mathbf{y}-\boldsymbol{\alpha})^\top\mathbf{H}\mathbf{u}\right] + \text{const.} \\
&= -\frac{\mathrm{E}\,\psi}{2}\Big[\xi^2\,\mathrm{tr}\left(\mathbf{H}^2\,\mathrm{E}[\mathbf{u}\mathbf{u}^\top]\right) - 2\xi(\mathbf{y}-\boldsymbol{\alpha})^\top\mathbf{H}\,\mathrm{E}\,\mathbf{u}\Big] + \text{const.}
\end{aligned}
$$

By completing the square, we get that $\tilde{q}(\xi) \equiv \mathrm{N}\left(d/c, (c\,\mathrm{E}[\psi])^{-1}\right)$, where

$$
c = \mathrm{tr}\left(\mathbf{H}^2\,\mathrm{E}[\mathbf{u}\mathbf{u}^\top]\right) \quad \text{and} \quad d = (\mathbf{y}-\boldsymbol{\alpha})^\top\mathbf{H}\,\mathrm{E}\,\mathbf{u}
$$

1.3.3 Distribution of $\tilde{q}(\psi)$

$$
\begin{aligned}
\log \tilde{q}(\psi) &= \mathrm{E}_{\mathbf{u},\xi}\left[\frac{n}{2}\log\psi - \frac{\psi}{2}\|\mathbf{y}-\boldsymbol{\eta}\|^2 + \frac{n}{2}\log\psi - \frac{\psi}{2}\|\mathbf{u}\|^2\right] + \text{const.} \\
&= n\log\psi - \mathrm{E}_{\mathbf{u},\xi}\left[\frac{\psi}{2}\left(\|\mathbf{y}-\boldsymbol{\eta}\|^2 + \|\mathbf{u}\|^2\right)\right] + \text{const.} \\
&= (n+1-1)\log\psi \\
&\quad - \psi \cdot \left(\frac{1}{2}\sum_{i=1}^{n}(y_i-\alpha)^2 + \frac{1}{2}\operatorname{tr}\left((\mathrm{E}[\xi^2]\mathbf{H}^2 + \mathbf{I}_n)\widetilde{\mathbf{U}}\right) - \mathrm{E}[\xi](\mathbf{y}-\boldsymbol{\alpha})^\top\mathbf{H}\,\mathrm{E}[\mathbf{u}]\right) + \text{const.}
\end{aligned}
$$

This is a gamma distribution with shape $s = n+1$ and rate $r = \frac{1}{2}\sum_{i=1}^{n}(y_i-\alpha)^2 + \frac{1}{2}\operatorname{tr}\left((\mathrm{E}[\xi^2]\mathbf{H}^2 + \mathbf{I}_n)\widetilde{\mathbf{U}}\right) - \mathrm{E}[\xi](\mathbf{y}-\boldsymbol{\alpha})^\top\mathbf{H}\,\mathrm{E}[\mathbf{u}]$. The mean is s/r, and the mean of $\log\psi$ is $\Psi(s) - \log r$, where Ψ is the digamma function.

1.4 Monitoring the lower bound

A convergence criterion would be when there is no more significant increase in the lower bound $\mathcal{L}$, as defined by

$$
\begin{aligned}
\mathcal{L} &= \int q(\mathbf{u},\xi,\psi)\log\left[\frac{p(\mathbf{y},\mathbf{u},\xi,\psi)}{q(\mathbf{u},\xi,\psi)}\right]\mathrm{d}\mathbf{u}\,\mathrm{d}\xi\,\mathrm{d}\psi \\
&= \mathrm{E}[\log p(\mathbf{y},\mathbf{u},\xi,\psi)] - \mathrm{E}[\log q(\mathbf{u},\xi,\psi)] \\
&= \mathrm{E}\left[\log p(\mathbf{y}|\boldsymbol{\eta})\right] + \mathrm{E}\left[\log p(\mathbf{u})\right] + \cancel{\mathrm{E}\left[\log p(\xi)\right]} + \cancel{\mathrm{E}\left[\log p(\psi)\right]} \\
&\quad - \mathrm{E}\left[\log q(\mathbf{u})\right] - \mathrm{E}\left[\log q(\xi)\right] - \mathrm{E}\left[\log q(\psi)\right]
\end{aligned}
$$

Definition 1 (Differential entropy)**.** *The differential entropy $\mathcal{H}$ of a pdf $p(x)$ is given by*

$$
\mathcal{H}(p) = -\int p(x)\log p(x)\,\mathrm{d}x = -\,\mathrm{E}_p[\log p(x)].
$$

Lemma 1. *Let $p(x)$ be the pdf of a random variable x. Then if*

(i) p is a univariate normal distribution with mean μ and variance ψ^{-1},

$$
\mathcal{H}(p) = \frac{1}{2}(1+\log 2\pi) - \frac{1}{2}\log\psi
$$

(ii) p is a d-dimensional normal distribution with mean μ and variance Σ,

$$
\mathcal{H}(p) = \frac{d}{2}(1+\log 2\pi) + \frac{1}{2}\log|\Sigma|
$$

(iii) p is the pdf of a gamma distribution with shape s and rate r,

$$\mathcal{H}(p) = s - \log r + \log \Gamma(s) + (1-s)\Psi(s)$$

where $\Psi(s) = \mathrm{d} \log \Gamma(s) / \, \mathrm{d}s = \Gamma'(s)/\Gamma(s)$ is the digamma function.

1.4.1 Terms involving distributions of $\mathbf{u}$

$$\begin{aligned}
\mathrm{E}\left[\log p(\mathbf{y}|\boldsymbol{\eta})\right] + \mathrm{E}\left[\log p(\mathbf{u})\right] - \mathrm{E}\left[\log q(\mathbf{u})\right] &= \mathrm{E}\left[-\frac{n}{2}\log 2\pi + \frac{n}{2}\log\psi - \frac{\psi}{2}\|\mathbf{y} - \boldsymbol{\alpha} - \xi\mathbf{H}\mathbf{u}\|^2\right] \\
&\quad + \mathrm{E}\left[-\frac{n}{2}\log 2\pi + \frac{n}{2}\log\psi - \frac{\psi}{2}\|\mathbf{u}\|^2\right] + \mathcal{H}(q(\mathbf{u})) \\
&= \cancel{-\frac{n}{2}\log 2\pi} + \frac{n}{2}\mathrm{E}[\log\psi] - \frac{\mathrm{E}\,\psi}{2}\mathrm{E}\left[\|\mathbf{y} - \boldsymbol{\alpha} - \xi\mathbf{H}\mathbf{u}\|^2\right] \\
&\quad - \frac{n}{2}\log 2\pi + \frac{n}{2}\mathrm{E}[\log\psi] - \frac{\mathrm{E}\,\psi}{2}\mathrm{E}[\|\mathbf{u}\|^2] \\
&\quad + \frac{n}{2}(1 + \cancel{\log 2\pi}) - \frac{1}{2}\log|\mathbf{A}| - \frac{n}{2}\log\mathrm{E}\,\psi \\
&= \frac{n}{2}\left(1 + 2\,\mathrm{E}[\log\psi] - \log\mathrm{E}\,\psi - \log 2\pi\right) - r\,\mathrm{E}\,\psi - \frac{1}{2}\log|\mathbf{A}| \\
&= \frac{n}{2}\left(1 + 2(\Psi(s) - \log r) - \log(s/r) - \log 2\pi\right) - r(s/r) - \frac{1}{2}\log|\mathbf{A}| \\
&= \frac{n}{2}\left(1 + 2\Psi(n+1) - \log r - \log(n+1) - \log 2\pi\right) - \frac{1}{2}\log|\mathbf{A}| - (n+1)
\end{aligned}$$

1.4.2 Terms involving distribution of $q(\xi)$

$$\begin{aligned}
-\mathrm{E}\left[\log q(\xi)\right] &= \mathcal{H}\big(q(\xi)\big) \\
&= \frac{1}{2}(1 + \log 2\pi) - \frac{1}{2}\log\operatorname{tr}\left(\mathbf{H}^2\widetilde{\mathbf{U}}\right) - \frac{1}{2}\log\mathrm{E}[\psi]
\end{aligned}$$

1.4.3 Terms involving distribution of $q(\psi)$

$$\begin{aligned}
-\mathrm{E}\left[\log q(\alpha)\right] &= \mathcal{H}\big(q(\psi)\big) \\
&= (n+1) - \log r + \log\Gamma(n+1) - n\Psi(n+1)
\end{aligned}$$

1.5 The variational Bayes EM algorithm

Algorithm 1 VB-EM algorithm for the probit I-prior model

procedure INITIALISE
  $\hat{\alpha} \leftarrow \sum_{i=1}^n y_i/n$
  $\tilde{\xi}^{(0)} \leftarrow 1$
  $\tilde{\xi}^{sq(0)} \leftarrow 1$ ▷ this is $\text{E}[\xi^2]$
  $\tilde{\psi}^{(0)} \leftarrow 1$
  $\tilde{\mathbf{u}}^{(0)} \leftarrow \mathbf{0}_n$ ▷ or draw $u_i^{(0)} \sim \text{N}(0,1)$ for $i = 1, \dots, n$.
end procedure

procedure UPDATE FOR $\mathbf{u}$ (time t)
  $\mathbf{A} \leftarrow \tilde{\xi}^{sq(t)}\mathbf{H}^2 + \mathbf{I}_n$
  $\mathbf{a} \leftarrow \tilde{\xi}^{(t)}\mathbf{H}(\mathbf{y} - \hat{\boldsymbol{\alpha}})$
  $\tilde{\mathbf{u}}^{(t+1)} \leftarrow \mathbf{A}^{-1}\mathbf{a}$
  $\widetilde{\mathbf{U}}^{(t+1)} \leftarrow \mathbf{A}^{-1}\big((1/\tilde{\psi}^{(t)})\mathbf{I}_n + \mathbf{a}\mathbf{a}^\top\mathbf{A}^{-1}\big)$
  $\text{logdetA}^{(t+1)} \leftarrow \log|\mathbf{A}|$
end procedure

procedure UPDATE FOR ξ (time t)
  $c \leftarrow c^{(t+1)} \leftarrow \text{tr}\left(\mathbf{H}^2\widetilde{\mathbf{U}}^{(t+1)}\right)$
  $d \leftarrow (\mathbf{y} - \hat{\boldsymbol{\alpha}})^\top\mathbf{H}\tilde{\mathbf{u}}^{(t+1)}$
  $\tilde{\xi}^{(t+1)} \leftarrow d/c$
  $\tilde{\xi}^{sq(t+1)} \leftarrow 1/(\tilde{\psi}^{(t)}c) + (d/c)^2$
end procedure

procedure UPDATE FOR ψ (time t)
  $r \leftarrow r^{(t+1)} \leftarrow \frac{1}{2}\sum_{i=1}^n (y_i - \hat{\alpha})^2 + \frac{1}{2}\text{tr}\left((\tilde{\xi}^{sq(t+1)}\mathbf{H}^2 + \mathbf{I}_n)\widetilde{\mathbf{U}}^{(t+1)}\right) - \tilde{\xi}^{(t+1)}(\mathbf{y} - \hat{\boldsymbol{\alpha}})^\top\mathbf{H}\tilde{\mathbf{u}}^{(t+1)}$
  $\tilde{\psi}^{(t+1)} \leftarrow (n+1)/r$
end procedure

procedure CALCULATE LOWER BOUND (time t)
  $\mathcal{L}^{(t)} \leftarrow \frac{n+1}{2}\big(1 - \log r^{(t+1)} - \log(n+1)\big) - \frac{n-1}{2}\log 2\pi + \log\Gamma(n+1) - \frac{1}{2}\left(\text{logdetA}^{(t+1)} + \log c^{(t+1)}\right)$
end procedure

procedure THE VB-EM ALGORITHM
  $t \leftarrow 0$
  while $\mathcal{L}^{(t+1)} - \mathcal{L}^{(t)} > \delta$ **or** $t < t_{max}$ **do**
    call UPDATE FOR $\mathbf{u}$
    call UPDATE FOR ξ
    call UPDATE FOR ψ
    call CALCULATE LOWER BOUND
    $t \leftarrow t + 1$
  end while
end procedure

return $(\hat{\mathbf{w}}, \hat{\alpha}, \hat{\lambda}, \hat{\psi}) \leftarrow (\tilde{\mathbf{u}}^{(t)}/\tilde{\psi}^{2(t)}, \hat{\alpha}, \tilde{\xi}^{(t)}/\tilde{\psi}^{2(t)}, \tilde{\psi}^{-2(t)})$ ▷ converged parameter estimates

return $(\hat{y}_1, \ldots, \hat{y}_n) \leftarrow \hat{\alpha}\mathbf{1} + \hat{\lambda}\mathbf{H}\hat{\mathbf{w}}$ ▷ fitted values