Incorporating Random effect into GP Sparse Approximation

Daeyoung Lim*
Department of Statistics
Korea University

February 13, 2016

1 GP linear mixed model

1.1 Model specifications

The model specifications remain the same except that the random effect term has been added.

$$\begin{aligned} y|\theta &\sim \mathcal{N}\left(Z\alpha + A\beta, \gamma^2 I_n\right), (n \times n) \\ \alpha|\sigma^2 &\sim \mathcal{N}\left(0, \frac{\sigma^2}{m} I_{2m}\right), (2m \times 1) \\ \beta &\sim \mathcal{N}\left(0, \Sigma_\beta\right), (s \times 1) \\ \lambda &\sim \mathcal{N}\left(\mu_\lambda, \Sigma_\lambda\right), (d \times 1) \\ \sigma &\sim \text{half-Cauchy}\left(A_\sigma\right) \\ \gamma &\sim \text{half-Cauchy}\left(A_\gamma\right) \end{aligned}$$

where A is the design matrix for the random effects and β is the parameter vector of the random effects.

1.2 Lower bound

$$p\left(y, \theta\right) = \mathcal{N}\left(y|Z\alpha + A\beta, \gamma^2 I_n\right) \mathcal{N}\left(\alpha \middle| 0, \frac{\sigma^2}{m} I_{2m}\right) \mathcal{N}\left(\beta|0, \Sigma_\beta\right) \mathcal{N}\left(\lambda, \mu_\lambda, \Sigma_\lambda\right) \text{HC}\left(A_\sigma\right) \text{HC}\left(A_\gamma\right)$$

*Prof. Taeryon Choi

$$
\begin{aligned}
(1)\ \mathbb{E}\left[\log p\left(y|\theta\right)\right] &= -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\mathbb{E}\left[\log\gamma^2\right] \\
&\quad - \frac{1}{2}\mathbb{E}\left[\frac{1}{\gamma^2}\right]\left\{y^T y - 2\left(\mathbb{E}\left[Z\right]m_\alpha + Am_\beta\right)^T y + \mathrm{Tr}\left(\mathbb{E}\left[Z^T Z\right]S_\alpha\right) + m_\alpha^T\mathbb{E}\left[Z^T Z\right]m_\alpha\right. \\
&\quad \left. + 2m_\beta^T A^T\mathbb{E}\left[Z\right]m_\alpha + \mathrm{Tr}\left(A^T A S_\beta\right) + m_\beta^T A^T A m_\beta\right\} \\
(2)\ \mathbb{E}\left[\log p\left(\alpha|\sigma\right)\right] &= -m\log\left(2\pi\right) - m\mathbb{E}\left[\log\sigma^2\right] + m\log m - \frac{m}{2}\mathbb{E}\left[\frac{1}{\sigma^2}\right]\mathrm{Tr}\left(S_\alpha + m_\alpha m_\alpha^T\right) \\
(3)\ \mathbb{E}\left[\log p\left(\beta\right)\right] &= -\frac{s}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left|\Sigma_\beta\right| - \frac{1}{2}\left\{\mathrm{Tr}\left(\Sigma_\beta^{-1}S_\beta\right) + m_\beta^T\Sigma_\beta^{-1}m_\beta\right\} \\
(4)\ \mathbb{E}\left[\log p\left(\lambda\right)\right] &= -\frac{d}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left|\Sigma_\lambda\right| - \frac{1}{2}\left(m_\lambda - \mu_\lambda\right)^T\Sigma_\lambda^{-1}\left(m_\lambda - \mu_\lambda\right) - \frac{1}{2}\mathrm{Tr}\left(\Sigma_\lambda^{-1}S_\lambda\right) \\
(5)\ \mathbb{E}\left[\log p\left(\sigma\right)\right] &= \log\left(2A_\sigma\right) - \log\pi - \mathbb{E}\left[\log\left(A_\sigma^2 + \sigma^2\right)\right] \\
(6)\ \mathbb{E}\left[\log p\left(\gamma\right)\right] &= \log\left(2A_\gamma\right) - \log\pi - \mathbb{E}\left[\log\left(A_\gamma^2 + \gamma^2\right)\right] \\
(1)\ \mathbb{E}\left[\log q\left(\alpha\right)\right] &= -m\log\left(2\pi\right) - \frac{1}{2}\log\left|S_\alpha\right| - m \\
(2)\ \mathbb{E}\left[\log q\left(\lambda\right)\right] &= -\frac{d}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left|S_\lambda\right| - \frac{d}{2} \\
(3)\ \mathbb{E}\left[\log q\left(\beta\right)\right] &= -\frac{s}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left|S_\beta\right| - \frac{s}{2} \\
(4)\ \mathbb{E}\left[\log q\left(\sigma\right)\right] &= -C_\sigma\frac{\mathcal{H}\left(2m, C_\sigma, A_\sigma^2\right)}{\mathcal{H}\left(2m-2, C_\sigma, A_\sigma^2\right)} - \log\mathcal{H}\left(2m-2, C_\sigma, A_\sigma^2\right) - 2m\mathbb{E}\left[\log\sigma\right] - \mathbb{E}\left[\log\left(A_\sigma^2 + \sigma^2\right)\right] \\
(5)\ \mathbb{E}\left[\log q\left(\gamma\right)\right] &= -C_\gamma\frac{\mathcal{H}\left(n, C_\gamma, A_\gamma^2\right)}{\mathcal{H}\left(n-2, C_\gamma, A_\gamma^2\right)} - \log\mathcal{H}\left(n-2, C_\gamma, A_\gamma^2\right) - n\mathbb{E}\left[\log\gamma\right] - \mathbb{E}\left[\log\left(A_\gamma^2 + \gamma^2\right)\right]
\end{aligned}
$$

$$
\mathcal{L} = (1) + (2) + (3) + (4) + (5) + (6) - (1) - (2) - (3) - (4) - (5)
$$

2 GP Logistic Model

2.1 Model specifications

For logistic models, we first postulate a link function, $g\left(\cdot\right)$ for the predictors.

$$
\begin{aligned}
y &= g^{-1}\left(\eta\right) + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \gamma^2 I_n\right) \\
g\left(\mathbb{E}\left[y\right]\right) &= \eta \\
\eta &= f\left(x\right) + A\beta \\
g^{-1}\left(x\right) &= e^x/\left(1+e^x\right) \\
f &\sim \mathcal{GP}\left(m\left(\cdot\right), \kappa\left(\cdot,\cdot\right)\right)
\end{aligned}
$$

Since y is a Bernoulli random variable, $\mathbb{E}\left[y\right] = \mathbb{P}\left(y=1\right)$. Without loss of generality, we will assume the mean function to be zero and that the Gaussian process has a sparse approximation representation as in Tan & Nott. Therefore,

$$
f\left(x\right) \approx \sum_{r=1}^{m}\left[a_r\cos\left\{\left(s_r \odot x\right)^T\lambda\right\} + b_r\sin\left\{\left(s_r \odot x\right)^T\lambda\right\}\right]
$$

and this could further be represented in matrix form which reduces this to a linear model.

$$
f\left(x\right) = Z\alpha
$$

$$
\begin{aligned}
\eta &= Z\alpha + A\beta \\
y &= \exp\left\{(Z\alpha + A\beta) - \log\left(\mathbf{1} + \exp\left\{Z\alpha + A\beta\right\}\right)\right\} + \epsilon
\end{aligned}
$$

Every scalar function applied to a vector or a matrix is done so elementwise. We think of the following priors:

$$
\begin{aligned}
\alpha|\sigma &\sim \mathcal{N}\left(0, \frac{\sigma^2}{m} I_{2m}\right) \\
\beta &\sim \mathcal{N}\left(\mu_\beta, \Sigma_\beta\right) \\
\lambda &\sim \mathcal{N}\left(\mu_\lambda, \Sigma_\lambda\right) \\
\sigma &\sim \text{half-Cauchy}\left(A_\sigma\right) \\
\gamma &\sim \text{half-Cauchy}\left(A_\gamma\right) \\
\theta &= (\alpha, \beta, \lambda, \sigma, \gamma)
\end{aligned}
$$

$$
\begin{aligned}
\log p\left(y, \theta\right) = {} & y^T\left(Z\alpha + A\beta\right) - \mathbf{1}_n^T \log\left(\mathbf{1}_n + \exp\left\{Z\alpha + A\beta\right\}\right) - \left(m + \frac{s+d}{2}\right)\log\left(2\pi\right) - m\log\sigma^2 + m\log m \\
& - \frac{m}{2\sigma^2}\alpha^T\alpha - \frac{1}{2}\log\left|\Sigma_\beta\right| - \frac{1}{2}\left(\beta - \mu_\beta\right)^T \Sigma_\beta^{-1}\left(\beta - \mu_\beta\right) - \frac{1}{2}\log\left|\Sigma_\lambda\right| - \frac{1}{2}\left(\lambda - \mu_\lambda\right)^T \Sigma_\lambda^{-1}\left(\lambda - \mu_\lambda\right) \\
& + \log\left(2A_\sigma\right) + \log\left(2A_\gamma\right) 2\log\pi - \log\left(A_\sigma^2 + \sigma^2\right) - \log\left(A_\gamma^2 + \gamma^2\right)
\end{aligned}
$$

Because $-\mathbf{1}_n^T \log\left(\mathbf{1}_n + \exp\left\{Z\alpha + A\beta\right\}\right)$ is analytically intractable for expectation which is essentially integration, we come up with the following approximation:

$$
\begin{aligned}
-\log\left(1 + e^x\right) &= \max_{\xi \in \mathbb{R}} \left\{B\left(\xi\right)x^2 - \frac{1}{2}x + C\left(\xi\right)\right\}, \quad \forall x \in \mathbb{R} \\
B\left(\xi\right) &= -\tanh\left(\xi/2\right)/\left(4\xi\right) \\
C\left(\xi\right) &= \xi/2 - \log\left(1 + e^\xi\right) + \xi\tanh\left(\xi/2\right)/4
\end{aligned}
$$

then

$$
\begin{aligned}
-\mathbf{1}_n^T \log\left\{\mathbf{1}_n^T + \exp\left(Z\alpha + A\beta\right)\right\} &\geq \mathbf{1}_n^T \left\{B\left(\xi\right) \odot \left(Z\alpha + A\beta\right)^2 - \frac{1}{2}\left(Z\alpha + A\beta\right) + C\left(\xi\right)\right\} \\
&= \left(Z\alpha + A\beta\right)^T \text{Dg}\left\{B\left(\xi\right)\right\}\left(Z\alpha + A\beta\right) - \frac{1}{2}\mathbf{1}_n^T\left(Z\alpha + A\beta\right) + \mathbf{1}_n^T C\left(\xi\right),
\end{aligned}
$$

where $\xi = (\xi_1, \ldots, \xi_n)$.

$$
\begin{aligned}
\log \underline{p}\left(y, \theta; \xi\right) = {} & y^T\left(Z\alpha + A\beta\right) + \left(Z\alpha + A\beta\right)^T \text{Dg}\left\{B\left(\xi\right)\right\}\left(Z\alpha + A\beta\right) - \frac{1}{2}\mathbf{1}_n^T\left(Z\alpha + A\beta\right) + \mathbf{1}_n^T C\left(\xi\right) \\
& - \left(m + \frac{s+d}{2}\right)\log\left(2\pi\right) - m\log\sigma^2 + m\log m - \frac{m}{2\sigma^2}\alpha^T\alpha - \frac{1}{2}\log\left|\Sigma_\beta\right| - \frac{1}{2}\left(\beta - \mu_\beta\right)^T \Sigma_\beta^{-1}\left(\beta - \mu_\beta\right) \\
& - \frac{1}{2}\log\left|\Sigma_\lambda\right| - \frac{1}{2}\left(\lambda - \mu_\lambda\right)^T \Sigma_\lambda^{-1}\left(\lambda - \mu_\lambda\right) + \log\left(2A_\sigma\right) + \log\left(2A_\gamma\right) 2\log\pi - \log\left(A_\sigma^2 + \sigma^2\right) \\
& - \log\left(A_\gamma^2 + \gamma^2\right)
\end{aligned}
$$

3 Sparse GP Probit Regression

3.1 Bayesian probit regression without random effects

The ordinary formulation goes like this:

$$\mathbb{P}\left(y_i = 1 | x_i, \beta\right) = \Phi\left(x_i^T \beta\right),$$

where Φ is the cumulative distribution of standard Gaussian distribution. Since the following likelihood

$$\mathbb{P}\left(Y = y | X, \beta\right) = \prod_{i=1}^{n} \Phi\left(x_i^T \beta\right)^{y_i} \left[1 - \Phi\left(x_i^T \beta\right)\right]^{1-y_i}$$

hinders the tractable calculation of the posterior, we devise the following latent variable:

$$Z_i = x_i^T \beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\left(0, 1\right)$$

for $i = 1, \ldots, n$ and let

$$y_i = \begin{cases} 1, & \text{if } Z_i \geq 0 \\ 0, & \text{otherwise} \end{cases}.$$

It automatically follows that

$$Z_i | y_i = 0, x_i, \beta \sim \mathcal{N}\left(x_i^T \beta, 1\right) \mathbf{1}\left[Z_i < 0\right]$$
$$Z_i | y_i = 1, x_i, \beta \sim \mathcal{N}\left(x_i^T \beta, 1\right) \mathbf{1}\left[Z_i \geq 0\right]$$

suggesting a truncated Gaussian for each Z_i depending on what value y_i takes on. According to probit regression without random effects, the introduction of latent variables enables the tractable computation of the posterior with regards to the parameters.

$$\pi\left(Z, y, \beta\right) = C\pi\left(\beta\right) \prod_{i=1}^{n} \left\{1\left[Z_i \geq 0\right] 1\left[y_i = 1\right] + 1\left[Z_i < 0\right] 1\left[y_i = 0\right]\right\} \phi\left(Z_i - x_i^T \beta\right)$$

where C is the proportionality constant and ϕ is the density of standard Gaussian. Following the mean-field approximation and computing the optimal density for Z_i, it becomes a truncated normal about zero with mean $x_i^T \beta$ and variance 1.

3.2 Incorporating random effects and nonparametric statistics

Taking one step forward, the random effects can be taken into consideration in many situations. Furthermore, we often do not know exactly what functional form the parameters should take on. Such situations greatly call for the use of nonparametric statistics.

$$\mathbb{P}\left(y_i = 1 | f\right) = \Phi\left(f\left(x_i\right)\right).$$

In the previous section where linear probit regression was considered, the functional form was assumed to be linear, i.e. $x_i^T \beta$. However, we now insist that f could be any function. Adopting Gaussion process prior for f and using sparse approximation,

$$\mathbb{P}\left(y_i = 1 | f\right) = \Phi\left(Z_i^T \alpha\right).$$

Note that Z_i is different from the one we used in the previous section. Moreover, we will assume there exist additive random effects:

$$\mathbb{P}\left(y_i = 1 | f, \alpha, \beta\right) = \Phi\left(z_i^T \alpha + a_i^T \beta\right).$$

Assume latent variables

$$y_i^* = z_i^T \alpha + a_i^T \beta + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, 1\right).$$

Let

$$y_i = \begin{cases} 1, & \text{if } y_i^* \geq 0 \\ 0, & \text{if } y_i^* < 0. \end{cases}$$

Let's assume priors

$$\begin{aligned} \alpha | \sigma &\sim \mathcal{N}\left(0, \frac{\sigma^2}{m} I_{2m}\right) \\ \beta &\sim \mathcal{N}\left(\mu_\beta, \Sigma_\beta\right) \\ \lambda &\sim \mathcal{N}\left(\mu_\lambda, \Sigma_\lambda\right) \\ \sigma &\sim \text{half-Cauchy}\left(A_\sigma\right) \\ \theta &= (\alpha, \beta, \lambda, \sigma) \end{aligned}$$

$$\pi\left(y, y^*, \theta\right) = C\pi\left(\alpha | \sigma\right)\pi\left(\beta\right)\pi\left(\lambda\right)\pi\left(\sigma\right)\prod_{i=1}^{n}\left\{1\left[y_i^* \geq 0\right]1\left[y_i = 1\right] + 1\left[y_i^* < 0\right]1\left[y_i = 0\right]\right\}\phi\left(y_i^* - z_i^T\alpha_i - a_i^T\beta\right)$$

3.2.1 $q\left(\alpha\right)$

Let $y^* = \begin{bmatrix} y_1^* & \dots & y_n^* \end{bmatrix}^T$.

$$\begin{aligned} \log q\left(\alpha\right) &\propto \left\langle \log \pi\left(\alpha | \sigma\right) \prod_{i=1}^{n} \phi\left(y_i^* - z_i^T\alpha - a_i^T\beta\right) \right\rangle \\ &\propto -\frac{m}{2}\left\langle \frac{1}{\sigma^2} \right\rangle \alpha^T\alpha + \left\langle \sum_{i=1}^{n}\left[-\frac{1}{2}\left\{y_i^{*2} - 2\left(z_i^T\alpha + a_i^T\beta\right)y_i^* + \left(z_i^T\alpha + a_i^T\beta\right)^2\right\}\right] \right\rangle \\ &\propto -\frac{m}{2}\left\langle \frac{1}{\sigma^2} \right\rangle \alpha^T\alpha + \left\langle -\frac{1}{2}\sum_{i=1}^{n}\left[-2y_i^* z_i^T\alpha + \alpha^T z_i z_i^T \alpha + 2\beta^T a_i z_i^T \alpha\right] \right\rangle \\ &\propto -\frac{m}{2}\left\langle \frac{1}{\sigma^2} \right\rangle \alpha^T\alpha - \frac{1}{2}\left\langle -2y^{*T} Z\alpha + \alpha^T\left(\sum_{i=1}^{n} z_i z_i^T\right)\alpha + 2\beta^T\left(\sum_{i=1}^{n} a_i z_i^T\right)\alpha \right\rangle \\ &\propto -\frac{m}{2}\left\langle \frac{1}{\sigma^2} \right\rangle \alpha^T\alpha - \frac{1}{2}\left\{-2\left\langle y^{*T} \right\rangle \left\langle Z \right\rangle \alpha + \alpha^T\left\langle Z^T Z \right\rangle \alpha + 2\left\langle \beta \right\rangle^T A^T \left\langle Z \right\rangle \alpha\right\} \\ &\propto -\frac{1}{2}\left\{\alpha^T\left(m\left\langle \frac{1}{\sigma^2} \right\rangle I_{2m} + \left\langle Z^T Z \right\rangle\right)\alpha - 2\left(\left\langle Z \right\rangle^T \left\langle y^* \right\rangle - \left\langle Z \right\rangle^T A^T \left\langle \beta \right\rangle\right)^T \alpha\right\} \\ q\left(\alpha\right) &= \mathcal{N}\left(\mu_{q(\alpha)}, \Sigma_{q(\alpha)}\right) \\ \mu_{q(\alpha)} &= \Sigma_{q(\alpha)}\left(\left\langle Z \right\rangle^T \left\langle y^* \right\rangle - \left\langle Z \right\rangle^T A^T \left\langle \beta \right\rangle\right) \\ \Sigma_{q(\alpha)} &= \left(m\left\langle \frac{1}{\sigma^2} \right\rangle I_{2m} + \left\langle Z^T Z \right\rangle\right)^{-1} \end{aligned}$$

3.2.2 $q(\beta)$

$$
\begin{aligned}
\log q(\beta) &\propto \log \pi(\beta) \prod_{i=1}^{n} \phi\left(y_i^* - z_i^T\alpha - a_i^T\beta\right) \\
&\propto -\frac{s}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_\beta| - \frac{1}{2}(\beta-\mu_\beta)^T\Sigma_\beta^T(\beta-\mu_\beta) \\
&\quad + \left\langle \sum_{i=1}^{T}\left[-\frac{1}{2}\left(y_i^* - 2\left(z_i^T\alpha + a_i^T\beta\right)y_i^* + \left(z_i^T\alpha + a_i^T\beta\right)^2\right)\right]\right\rangle \\
&\propto -\frac{1}{2}\left(\beta^T\Sigma_\beta^{-1}\beta - 2\mu_\beta^T\Sigma_\beta^{-1}\beta\right) + \left\langle \sum_{i=1}^{n}\left[-\frac{1}{2}\left(-2y_i^* a_i^T\beta + 2\alpha^T z_i a_i^T\beta + \beta^T a_i a_i^T\beta\right)\right]\right\rangle \\
&\propto -\frac{1}{2}\left(\beta^T\Sigma_\beta^{-1}\beta - 2\mu_\beta^T\Sigma_\beta^{-1}\beta\right) - \frac{1}{2}\left(-2\left\langle y^{*T}\right\rangle A\beta + 2\left\langle\alpha^T\right\rangle\left\langle Z^T\right\rangle A\beta + \beta^T A^T A\beta\right) \\
&\propto -\frac{1}{2}\left\{\beta^T\Sigma_\beta^{-1}\beta + \beta^T A^T A\beta - 2\mu_\beta^T\Sigma_\beta^{-1}\beta + 2\left\langle\alpha^T\right\rangle\left\langle Z^T\right\rangle A\beta - 2\left\langle y^{*T}\right\rangle A\beta\right\} \\
&\propto -\frac{1}{2}\left\{\beta^T\left(\Sigma_\beta^{-1} + A^T A\right)\beta - 2\left(\Sigma_\beta^{-1}\mu_\beta + A^T\langle y^*\rangle - A^T\langle Z\rangle\langle\alpha\rangle\right)^T\beta\right\} \\
q(\beta) &= \mathcal{N}\left(\mu_{q(\beta)}, \Sigma_{q(\beta)}\right) \\
\mu_{q(\beta)} &= \Sigma_{q(\beta)}\left(\Sigma_\beta^{-1}\mu_\beta + A^T\langle y^*\rangle - A^T\langle Z\rangle\langle\alpha\rangle\right) \\
\Sigma_{q(\beta)} &= \left(\Sigma_\beta^{-1} + A^T A\right)^{-1}
\end{aligned}
$$

3.2.3 $q(y^*)$

$$
\begin{aligned}
\log q(y^*) &\propto \left\langle \log \prod_{i=1}^{n}\left\{1[y_i^* \geq 0]\,1[y_i = 1] + 1[y_i^* < 0]\,1[y_i = 0]\right\}\phi\left(y_i^* - z_i^T\alpha - a_i^T\beta\right)\right\rangle \\
&\propto \sum_{i=1}^{n}\log\left\{1[y_i^* \geq 0]\,1[y_i = 1] + 1[y_i^* < 0]\,1[y_i = 0]\right\} \\
&\quad + \left\langle \sum_{i=1}^{n}\left[-\frac{1}{2}\left(y_i^{*2} - 2\left(z_i^T\alpha + a_i^T\beta\right)y_i^* + \left(z_i^T\alpha + a_i^T\beta\right)^2\right)\right]\right\rangle \\
&\propto \sum_{i=1}^{n}\log\left\{1[y_i^* \geq 0]\,1[y_i = 1] + 1[y_i^* < 0]\,1[y_i = 0]\right\} \\
&\quad - \frac{1}{2}\left(y^{*T}y^* - 2\left(\langle Z\rangle\langle\alpha\rangle + A\langle\beta\rangle\right)^T y^*\right) \\
q(y^*) &= \mathcal{TN}\left(\langle Z\rangle\langle\alpha\rangle + A\langle\beta\rangle, I_n\right)
\end{aligned}
$$

where $\mathcal{TN}$ indicates truncated normal distribution, in this case multivariate. Each element of y^* is truncated at zero. The direction of truncation remains the same. Assuming $y_i = 1$,

$$
\mu_{q(y_i^*)} = \left\langle z_i^T\right\rangle\mu_{q(\alpha)} + a_i^T\mu_{q(\beta)} + \frac{\phi\left(\left\langle z_i^T\right\rangle\mu_{q(\alpha)} + a_i^T\mu_{q(\beta)}\right)}{\Phi\left(\left\langle z_i^T\right\rangle\mu_{q(\alpha)} + a_i^T\mu_{q(\beta)}\right)}
$$

3.2.4 $q(\sigma)$

$$\begin{aligned}
\log q(\sigma) &\propto \langle \log \pi(\alpha|\sigma)\, \pi(\sigma) \rangle \\
&\propto -m \log \sigma^2 - \frac{m}{2} \left\langle \alpha^T \alpha \right\rangle / \sigma^2 - \log\left(A_\sigma^2 + \sigma^2\right) \\
q(\sigma) &\propto \frac{\exp\left(-C_\sigma/\sigma^2\right)}{\sigma^{2m}\left(A_\sigma^2 + \sigma^2\right)} \\
C_\sigma &= \frac{m}{2}\left(\mathrm{Tr}\left(\Sigma_{q(\alpha)}\right) + \mu_{q(\alpha)}^T \mu_{q(\alpha)}\right)
\end{aligned}$$

3.2.5 LB: $\mathbb{E}[\log p(y, y^*|\theta)]$

Recall that

$$\pi(y, y^*, \theta) = C\pi(\alpha|\sigma)\, \pi(\beta)\, \pi(\lambda)\, \pi(\sigma) \prod_{i=1}^{n} \left\{1[y_i^* \geq 0]\, 1[y_i = 1] + 1[y_i^* < 0]\, 1[y_i = 0]\right\} \phi\left(y_i^* - z_i^T \alpha_i - a_i^T \beta\right)$$

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}[\log \pi(y, y^*, \theta)] - \mathbb{E}[\log q(y^*)\, q(\theta)] \\
&= \mathbb{E}[\log p(y|y^*, \theta)] + \mathbb{E}[\log p(y^*|\theta)] + \mathbb{E}[\log p(\theta)] - \mathbb{E}[\log q(\theta)]
\end{aligned}$$

The above equation includes all the nodes in the graphical model. Therefore, we compute the expecation of the logarithm of each node.

$$\begin{aligned}
\log p(y, y^*|\theta) &= \log \prod_{i=1}^{n} \left\{1[y_i^* \geq 0]\, 1[y_i = 1] + 1[y_i^* < 0]\, 1[y_i = 0]\right\} \phi\left(y_i^* - z_i^T \alpha - a_i^T \beta\right) \\
&\implies \mathbb{E}\left[\sum_{i=1}^{n} \log\left\{1[y_i^* \geq 0]\, 1[y_i = 1] + 1[y_i^* < 0]\, 1[y_i = 0]\right\}\right] + \sum_{i=1}^{n} \mathbb{E}\left[\log \phi\left(y_i^* - z_i^T \alpha - a_i^T \beta\right)\right]
\end{aligned}$$

If we set $U = \{1[y_i^* \geq 0]\, 1[y_i = 1] + 1[y_i^* < 0]\, 1[y_i = 0]\}$ and let it be a new random variable, it is always 1 with probability 1. Therefore, the logarithm of U, i.e. $\log U$ is always 0 which terminates the need for calculating the expected value since it's 0 regardless. We proceed to the remaining terms.

$$\sum_{i=1}^{n} \mathbb{E}\left[\log \phi\left(y_i^* - z_i^T \alpha - a_i^T \beta\right)\right] \qquad \cdots (1)$$

The entropy of the variational distribution of $q\left(y_i^*\right)$ should be coupled with eqn (1) to simplify the calculation.

$$
\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[\log \phi\left(y_i^* - z_i^T \alpha - a_i^T \beta\right) - \log \phi\left(y_i^* - \left\langle z_i^T \right\rangle \mu_{q(\alpha)} - a_i^T \mu_{q(\beta)}\right)\right] \\
&+ \sum_{i=1}^{n} \log\left(\left\{\Phi\left(\langle z_i \rangle^T \mu_{q(\alpha)} + a_i^T \mu_{q(\beta)}\right)\right\}^{y_i} \left\{1 - \Phi\left(\langle z_i \rangle^T \mu_{q(\alpha)} + a_i^T \mu_{q(\beta)}\right)\right\}^{1-y_i}\right) \\
&= \sum_{i=1}^{n} \mathbb{E}\left[-\frac{1}{2}\left(-2\left(z_i^T \alpha + a_i^T \beta\right) y_i^* + \left(z_i^T \alpha + a_i^T \beta\right)^2 + 2\left(\langle z_i \rangle^T \mu_{q(\alpha)} + a_i^T \mu_{q(\beta)}\right) y_i^*\right)\right] \\
&\quad + \frac{1}{2} \sum_{i=1}^{n} \left(\left(\langle z_i \rangle^T \mu_{q(\alpha)} + a_i^T \mu_{q(\beta)}\right)^2\right) \\
&\quad + \sum_{i=1}^{n} \log\left(\left\{\Phi\left(\langle z_i \rangle^T \mu_{q(\alpha)} + a_i^T \mu_{q(\beta)}\right)\right\}^{y_i} \left\{1 - \Phi\left(\langle z_i \rangle^T \mu_{q(\alpha)} + a_i^T \mu_{q(\beta)}\right)\right\}^{1-y_i}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{2}\left(\operatorname{Tr}\left(\left\langle Z^T Z \right\rangle \Sigma_{q(\alpha)}\right) + \mu_{q(\alpha)}^T \left(\left\langle Z^T Z \right\rangle - \langle Z \rangle^T \langle Z \rangle\right) \mu_{q(\alpha)} + \operatorname{Tr}\left(A^T A \Sigma_{q(\beta)}\right)\right) \\
&\quad + \sum_{i=1}^{n} \log\left(\left\{\Phi\left(\langle z_i \rangle^T \mu_{q(\alpha)} + a_i^T \mu_{q(\beta)}\right)\right\}^{y_i} \left\{1 - \Phi\left(\langle z_i \rangle^T \mu_{q(\alpha)} + a_i^T \mu_{q(\beta)}\right)\right\}^{1-y_i}\right)
\end{aligned}
$$