1 Stochastic Variational Inference in Linear Regression

Model $y = \mathbf{X}\beta + \mathbf{e}, \quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

- $\beta \sim \mathcal{N}(\mu_0, \Sigma_0)$
- $\sigma^2 \sim \text{InvGam}(A, B)$

Variational distributions

- $q(\beta) \sim \mathcal{N}\left(\mu_q, \Sigma_q\right)$
- $q\left(\sigma^2\right) \sim \text{InvGam}\left(A_q, B_q\right)$

Let $h(\theta) = p(y \mid \theta)\, p(\theta)$.

$$\log h(\theta) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta) - \frac{1}{2}\log|\Sigma_0| - \frac{p}{2}\log(2\pi) \tag{1}$$

$$- \frac{1}{2}(\beta - \mu_0)'\Sigma_0^{-1}(\beta - \mu_0) - (A+1)\log\sigma^2 - \frac{B}{\sigma^2} + A\log B - \log\Gamma(A) \tag{2}$$

$$\log q_\lambda(\theta) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log\left|\Sigma_q\right| - \frac{1}{2}\left(\beta - \mu_q\right)'\Sigma_q^{-1}\left(\beta - \mu_q\right) - \left(A_q + 1\right)\log\sigma^2 \tag{3}$$

$$- \frac{B_q}{\sigma^2} + A_q \log B_q - \log\Gamma\left(A_q\right) \tag{4}$$

1.1 Reparameterization

Let $\Sigma_q^{-1} = \Omega = CC'$ where C is the Cholesky factor lower triangular matrix of Ω. Then with $s \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_p\right)$, we can recast β as

$$\beta = C'^{-1}s + \mu_q \tag{5}$$

Now, the log-joint density becomes

$$\log h(\theta) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(y - XC'^{-1}s - X\mu_q\right)'\left(y - XC'^{-1}s - X\mu_q\right) - \frac{1}{2}\log|\Sigma_0| - \frac{p}{2}\log(2\pi) \tag{6}$$

$$- \frac{1}{2}\left(C'^{-1}s + \mu_q - \mu_0\right)'\Sigma_0^{-1}\left(C'^{-1}s + \mu_q - \mu_0\right) - (A+1)\log\sigma^2 - \frac{B}{\sigma^2} \tag{7}$$

$$+ A\log B - \log\Gamma(A) \tag{8}$$

$$\log q_\lambda(\theta) = -\frac{p}{2}\log(2\pi) + \log|C| - \frac{1}{2}s's - (A_q + 1)\log\sigma^2 \tag{9}$$

$$- \frac{B_q}{\sigma^2} + A_q \log B_q - \log\Gamma\left(A_q\right) \tag{10}$$

2 Multivariate Gaussian Distribution

We will assume that the variational posterior distribution $q_\lambda(\theta)$ is a multivariate Gaussian distribution where λ denotes the natural parameters in the exponential family representation of the density given by

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \Sigma^{-1}\mu \\ -1/2 D_d' \operatorname{vec}\left(\Sigma^{-1}\right) \end{bmatrix} \tag{11}$$

D_d is the duplication matrix of order d, which is the unique $d^2 \times d(d+1)/2$ matrix of zeros and ones such that $D_d \operatorname{vech}(A) = \operatorname{vec}(A)$ for symmetric $d \times d$ matrices A and its Moore-Penrose inverse is

$$D_d^\dagger = \left(D_d' D_d\right)^{-1} D_d' \tag{12}$$

On the contrary, L_d is the elimination matrix such that

$$L_d \operatorname{vec}(A) = \operatorname{vech}(A) \tag{13}$$

With these notations and natural parameters, the original parameters μ, Σ can be expressed as

$$\mu = -\frac{1}{2}\left\{\operatorname{vec}^{-1}\left(D_d^\dagger \lambda_2\right)\right\}^{-1} \lambda_1, \qquad \Sigma = -\frac{1}{2}\left\{\operatorname{vec}^{-1}\left(D_d^{\dagger\prime} \lambda_2\right)\right\}^{-1} \tag{14}$$

Then, the exponential family representation becomes

$$q_\lambda(\theta) = \exp\left(T(\theta)'\lambda - Z(\lambda)\right) \tag{15}$$

where $T(\theta)$ is the sufficient statistic

$$T(\theta) = \begin{bmatrix} \theta \\ \operatorname{vech}\left(\theta\theta'\right) \end{bmatrix} \tag{16}$$

The Fisher's information matrix $I_F(\lambda) = \operatorname{Cov}_\lambda(T(\theta))$ has an inverse

$$I_F(\lambda)^{-1} = \begin{bmatrix} \Sigma^{-1} + M' S^{-1} M & -M' S^{-1} \\ -S^{-1} M & S^{-1} \end{bmatrix} \tag{17}$$

where $M = 2D_d^\dagger(\mu \otimes I_d)$ and $S = 2D_d^\dagger(\Sigma \otimes \Sigma)D_d^{\dagger\prime}$. Now

$$\nabla_\lambda \log q_\lambda(\theta) = \begin{bmatrix} \theta - \mu \\ \text{vech}\left(\theta\theta' - \Sigma - \mu\mu'\right) \end{bmatrix} \tag{18}$$

3 Gradients

$$\nabla_{\mu_q} \log q_\lambda(\theta) = -\Sigma_q^{-1}\left(\mu_q - \beta\right) \tag{19}$$

$$\nabla_{\Sigma_q} \log q_\lambda(\theta) = -\frac{1}{2}\left(\Sigma_q^{-1} + \left(\beta - \mu_q\right)\left(\beta - \mu_q\right)'\right) \tag{20}$$

$$\nabla_{A_q} \log q_\lambda(\theta) = -\log \sigma^2 + \log B_q - \psi\left(A_q\right) \tag{21}$$

$$\nabla_{B_q} \log q_\lambda(\theta) = -\frac{1}{\sigma^2} + \frac{A_q}{B_q} \tag{22}$$

Lower bound, log-derivative trick

$$\nabla_\lambda \mathcal{L}(\lambda) = \nabla_\lambda \int \left(\log h(\theta) - \log q_\lambda(\theta)\right) q_\lambda(\theta)\, \theta \tag{23}$$

$$= \int \log h(\theta) \nabla_\lambda \log q_\lambda(\theta) q_\lambda(\theta)\, d\theta - \int \log q_\lambda(\theta) \nabla_\lambda \log q_\lambda)(\theta) q_\lambda(\theta)\, \theta \tag{24}$$

$$= \int \nabla_\lambda \log q_\lambda(\theta) \left(\log h(\theta) - \log q_\lambda(\theta)\right) q_\lambda(\theta)\, d\theta \tag{25}$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} \nabla_\lambda \log q_\lambda\left(\theta^{(s)}\right)\left(\log h(\theta^{(s)}) - \log q_\lambda(\theta^{(s)})\right) \tag{26}$$

where $\theta^{(s)} \sim q_\lambda(\theta),\ s = 1, \ldots, S$.

4 Step Size

For the step size sequence ρ_t, the Robbins-Monro approximation states that it should satisfy two conditions

- $\sum_{t=1}^{\infty} \rho_t = \infty$
- $\sum_{t=1}^{\infty} \rho_t^2 < \infty$

which the first indicates that the step size should be large enough that the stochastic search is able to sweep through all possible zones of the parameter space, and the second indicates that despite the first condition, the variation should not be so big that it is beyond control. The recommended Robbins-Monro sequence is

$$\rho = (t + \tau)^{-\kappa} \tag{27}$$

where $\kappa \in (0.5, 1]$, the forgetting rate, controls how quickly old information decays and $\tau \geq 0$, the delay, downweights early iterations. In many cases, τ is set to 1.

5 Inverse Gamma Distribution

The log-density of an inverse-gamma distribution is

$$A \log B - \log \Gamma(A) - (A + 1) \log \theta - B/\theta \tag{28}$$

Therefore for A, B, the gradient is given as

$$\nabla_\lambda q_\lambda(\theta) = \begin{bmatrix} \log B - \psi(A) - \log \theta \\ A/B - 1/\theta \end{bmatrix} \tag{29}$$

Now to get the Fisher's information matrix, we need the following

$$\begin{aligned}
\text{Var}(-\log \theta) &= \psi_1(A) && (30)\\
\text{Var}(-1/\theta) &= A/B^2 && (31)\\
\text{Cov}(-\log \theta, -1/\theta) &= \mathbf{E}\left(\frac{1}{\theta} \log \theta\right) - \mathbf{E}(\log \theta)\, \mathbf{E}\left(\frac{1}{\theta}\right) && (32)\\
&= -\mathbf{E}\left(\frac{1}{\theta} \log \frac{1}{\theta}\right) + \mathbf{E}\left(\log \frac{1}{\theta}\right) \mathbf{E}\left(\frac{1}{\theta}\right) && (33)\\
&= \frac{A}{B} \left(\psi(A) - \psi(A + 1)\right) && (34)\\
G_\theta &= \begin{bmatrix} \psi_1(A) & -\frac{1}{B} \\ -\frac{1}{B} & A/B^2 \end{bmatrix} && (35)
\end{aligned}$$

since $\psi(a + 1) = \psi(a) + 1/a$

6 Reparameterization

If $q(\beta) = \mathcal{N}\left(\mu_q, \Sigma_q\right)$ and if we let $\Sigma_q^{-1} = CC'$ and $\phi(z) = \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$,

- $\beta = C^{-1'}z + \mu_q$
- $dz = |C|d\beta$
- $q(\beta) = |C|\phi(z)$

are all true. Then

$$\int q(\beta) \log \frac{h(\theta)}{q(\beta)}\, d\beta = \int \phi(z) \log \frac{h\left(C^{-1'}z + \mu_q\right)}{\phi(z)|C|}\, dz \tag{36}$$

$$= \mathbf{E}_{\phi(z)}\left[h\left(C^{-1'}z + \mu_q\right)\right] - \log|C| - \mathcal{H}(z) \tag{37}$$