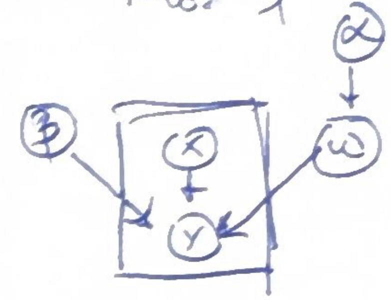
Start from beginning.

④ BAYESIAN LINEAR REGRESSION

VARIATIONAL INFERENCE

Model



Full joint distribution; $p(y, x, \omega, \alpha, \beta) = p(y|x, \omega, \beta)\, p(\omega|\alpha)\, p(\beta)\, p(\alpha)\, p(x)$

but $p(x)$ is a constant given data are observed.

$$p(y, x, \omega, \alpha, \beta) \propto p(y, \omega, \alpha, \beta) = p(y|\cancel{x}, \omega, \beta)\, p(\omega|\alpha)\, p(\alpha)\, p(\beta)$$

for $Z = \{\omega, \alpha, \beta\}$

$$\ln p(y) = \ln \frac{p(y, Z)}{p(Z|y)} \quad \text{since } p(y,Z) = p(Z|y)\,p(y)$$

using the derivation from pg 1

$$\ln p(y) = \ln \frac{p(y,Z)}{p(Z|y)} = \underbrace{\int q(Z) \ln \frac{p(y,Z)}{q(Z)}\, dZ}_{ELBO(q)} + \underbrace{\int q(Z) \ln \frac{q(Z)}{p(Z|y)}\, dZ}_{KL\, q\|p}$$

not symmetric.

$y = \omega^T x + \varepsilon$

$\Rightarrow p(y|x, \omega, \beta) \sim \mathcal{N}(y|X\omega, \beta^{-1} I)$

i.e. $\mu_y = X\omega$

$\sigma_y^2 = \beta^{-1} I$

$p(\omega|\alpha) \sim \mathcal{N}(\omega|0, \alpha^{-1} I)$

i.e. $\mu_\omega = 0$

$\sigma_\omega^2 = \alpha^{-1}$

$p(\alpha) \sim \Gamma(\alpha|a_0, b_0)$

$p(\beta) \sim \Gamma(\beta|c_0, d_0)$

The KL-divergence measures the difference b/w our posterior and our approximation.

∴ we want to minimize our KL-divergence. We can't do this directly b/c we don't know the ~~post~~ posterior :(

~~But minimizing the KL-div is eq (exactly) equal to maximizing the ELBO.~~

~~Given our log-joint $\ln p(y, \omega, \alpha, \beta) = \ln(y|x, \omega, \beta) + \ln p(\omega|\alpha) + \ln p(\alpha) + \ln p(\beta)$,~~

~~we're trying to find the best approximation~~

but note that $p(z|y) = \dfrac{p(y,z)}{p(y)}$

$\therefore$ we can re-write the $KL_{q||p}$ as, $\int q(z) \ln \dfrac{q(z)}{p(z|y)} dz = \int q(z) \ln \dfrac{q(z)}{\frac{p(y,z)}{p(y)}} dz$

$\Rightarrow \int q(z) \ln \dfrac{q(z)\, p(y)}{p(y,z)}$ ← constant w.r.t. z.

$\therefore$ minimizing KL-div w.r.t. posterior $\propto$ ← proportional.

" " " full-joint. !!!

$\equiv$ maximizing the ELBO

$q(z) = q(w, \alpha, \beta) \xrightarrow[\text{field assumption}]{\text{Mean}} q(w)\,q(\alpha)\,q(\beta) \longrightarrow$ same in model structure

~~Start w/ optimizing $q(w) \to q^*(w) = \mathbb{E}_{q(\alpha)q(\beta)}[\ln p(y,z)]$~~

$q(z) \approx q(w)\,q(\alpha)\,q(\beta) \approx p(y, w, \alpha, \beta)$

~~To find.~~ First ~~find~~ optimize $q(w) \to q^*(w)$. Instead of marginalizing, we can take the expectation.

Consider,

$$\int q(z) \ln \frac{q(z)}{p(z|y)} dz = \int q(w)\,q(\alpha)\,q(\beta) \ln \frac{q(w)\,q(\alpha)\,q(\beta)}{p(w,\alpha,\beta|y)} dw\, d\alpha\, d\beta.$$

$$\Rightarrow \int q(w)\,q(\alpha)\,q(\beta) \left[\ln q(w) + \ln q(\alpha) + \ln q(\beta) - \ln p(w,\alpha,\beta|y)\right] dw\, d\alpha\, d\beta$$

If we want to optimize wrt. $q(\omega)$, then $q(\alpha)$ and $q(\beta)$ are constants. There is some complex dependency in $p(\omega, \alpha, \beta \mid y)$ why we can "remove" by taking the expectation wrt. $q(\alpha), q(\beta)$

$$\therefore q^*(\omega) \propto \mathbb{E}_{q(\alpha)q(\beta)} \cancel{p(y \mid \omega} \; p(\omega, \alpha, \beta \mid y) \Rightarrow \mathbb{E}_{q(\alpha)q(\beta)}\left[\frac{p(y, \omega, \alpha, \beta)}{p(y)}\right]$$

$$\ln q^*(\omega) = \mathbb{E}_{q(\alpha)q(\beta)}[\ln p(y, \omega, \alpha, \beta) - \underbrace{\ln(y)}_{\text{constant}}]$$

$$= \mathbb{E}_{q(\alpha)q(\beta)}[\ln p(y \mid x, \omega, \beta) + \ln p(\omega \mid \alpha) + \underbrace{\ln p(\alpha) + \ln p(\beta)}_{\text{constant wrt } \omega.}]$$

Distribute expectations

$$= \mathbb{E}_{q(\beta)} \ln p(y \mid x, \omega, \beta) + \mathbb{E}_{q(\alpha)} \ln p(\omega \mid \alpha) + \underbrace{\mathbb{E}_{q(\alpha)} \ln p(\alpha) + \mathbb{E}_{q(\beta)} \ln p(\beta)}_{\text{constant wrt } \omega}$$

$$p(y \mid x, \omega, \beta) = \mathcal{N}(y \mid X\omega, \beta^{-1}\mathbb{I})$$

$$p(\omega \mid \alpha) = \mathcal{N}(\omega \mid 0, \alpha^{-1}\mathbb{I})$$

plug in ~~$x = \mu$~~ $x = \omega$, $\mu = 0$, $\Sigma = \alpha^{-1}\mathbb{I}$

$$\ln p(\omega \mid \alpha) = \ln\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)^T\right]\right]$$

$$= \ln\left[\frac{1}{(2\pi)^{d/2}|\alpha^{-1}\mathbb{I}|^{1/2}} \exp\left[-\frac{1}{2}(\omega-\mu)^T[\alpha^{-1}\mathbb{I}]^{-1}(\omega-\mu)^T\right]\right]$$

$\mu = 0$

$$\Rightarrow \underbrace{\ln\left[\frac{1}{(2\pi)^{d/2}|\alpha^{-1}\mathbb{I}|^{1/2}}\right]}_{\text{constant wrt } \omega.} - \frac{1}{2}(\omega-\mu)^T[\alpha^{-1}\mathbb{I}]^{-1}(\omega-\mu)^T \Rightarrow -\frac{1}{2}\omega^T \alpha \mathbb{I} \omega = \boxed{\frac{-\alpha}{2}\omega^T\omega}$$

Do same for likelihood

$$p(y|x,\omega,\beta) \sim \mathcal{N}(y|X\omega, \beta^{-1}\mathbb{I}) = \ln\left[\overbrace{\frac{1}{(2\pi)^{d/2}|\beta^{-1}\mathbb{I}|^{1/2}}}^{\text{constant wrt. } \omega} \exp\left(-\frac{1}{2}(y-X\omega)^T(\beta^{-1}\mathbb{I})^{-1}(y-X\omega)\right)\right]$$

$$\ln p(y|x,\omega,\beta) = \cancel{\ln\left[\underbrace{\frac{1}{(2\pi)^{d/2}|\beta^{-1}\mathbb{I}|^{1/2}}}_{\cancel{\text{Constant wrt. } \omega}} \exp\left(\frac{1}{2}(X\omega-\mu)^T(\beta^{-1}\mathbb{I})(X\omega-\mu)\right)\right]}$$

$$\cancel{\Rightarrow -\frac{1}{2}(X\omega-\mu)^T(\beta^{-1}\mathbb{I})(X\omega-\mu)}$$

$$\Rightarrow -\frac{1}{2}(y-X\omega)^T(\beta^{-1}\mathbb{I})^{-1}(y-X\omega)$$

$$\Rightarrow -\frac{\beta}{2}(y-X\omega)^T(y-X\omega)$$

Back to $q^*(\omega) = \mathbb{E}_{q(\beta)}\ln p(y|x,\omega,\beta) + \mathbb{E}_{q(\alpha)}\ln p(\omega|\alpha)$

$$\Rightarrow \mathbb{E}_{q(\beta)}\left[-\frac{\beta}{2}\underbrace{(y-X\omega)^T(y-X\omega)}\right] + \mathbb{E}_{q(\alpha)}\left[-\frac{\alpha}{2}\omega^T\omega\right]$$

$$y^Ty - 2\omega^TX^Ty + \omega^TX^TX\omega \qquad \cancel{\Rightarrow \mathbb{E}_{q(\beta)}\left[-\frac{\beta}{2}\right.}$$

* Everything is constant wrt. α & β except α and β

$$\therefore \Rightarrow -\frac{\mathbb{E}_{q(\beta)}(\beta)}{2}\left[y^Ty - 2\omega^TX^Ty + \omega^TX^TX\omega\right] - \frac{\mathbb{E}_{q(\alpha)}(\alpha)}{2}\omega^T\omega$$

$$\Rightarrow -\frac{\mathbb{E}_{q(\beta)}(\beta)}{2}y^Ty + \mathbb{E}_{q(\beta)}(\beta)\omega^TX^Ty \underbrace{- \frac{\mathbb{E}_{q(\beta)}(\beta)}{2}\omega^TX^TX\omega - \frac{\mathbb{E}_{q(\alpha)}(\alpha)}{2}\overbrace{\omega^T\omega}^{\omega^T\mathbb{I}\omega}}_{\omega^T\left[-\frac{\mathbb{E}_{q(\beta)}(\beta)}{2}X^TX\right]\omega + \omega^T\left[-\frac{\mathbb{E}_{q(\alpha)}(\alpha)}{2}\mathbb{I}\right]\omega}$$

constant — linear term — quadratic terms

$$\Rightarrow \underbrace{\frac{-\mathbb{E}_{q(\beta)}(\beta)}{2} y^T y}_{\text{constant}} + \underbrace{\mathbb{E}_{q(\beta)}(\beta)\, \omega^T X^T y}_{\text{linear term}} \underbrace{- \frac{1}{2}\omega^T \Big[\underbrace{\mathbb{E}_{q(\beta)}(\beta) X^T X + \mathbb{E}_{q(\alpha)}(\alpha)\, \mathbb{I}}_{\Sigma_\omega^{-1}}\Big]\omega}_{\text{quadratic terms}}$$

ω subscript.

All terms are constants, e.g.: $x^T \Sigma^{-1} \mu : (1\times n)(n \times n)(n \times 1)$ ∴ transpose is the same

Expansion of Gaussian:

$$-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) = -\frac{1}{2}\left[x\Sigma^{-1} - \mu\Sigma^{-1}\right]^T (x-\mu) = -\frac{1}{2}\Big[x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu - \underbrace{\mu^T\Sigma^{-1}x}_{\text{transpose}} + \mu^T\Sigma^{-1}\mu\Big]$$

~~grouping x~~

$$\Rightarrow -\frac{1}{2}\left[x\Sigma^{-1}x - x^T\Sigma^{-1}\mu - x^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu\right]$$

$$\Rightarrow -\frac{1}{2}\left[x\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu\right]$$

$\mu :=$ constant ∴ $\Rightarrow -\frac{1}{2}\Big[\underbrace{x\Sigma^{-1}x}_{\text{quadratic term}} - \underbrace{2x^T\Sigma^{-1}\mu}_{\text{linear term}}\Big]$ + constant $\Rightarrow x^T\Sigma^{-1}\mu - \frac{1}{2}x\Sigma^{-1}x$

$x \to y$

$\mu \to X\omega$

$$\boxed{\Sigma_\omega^{-1} = \mathbb{E}_{q(\beta)}(\beta) X^T X + \mathbb{E}_{q(\alpha)}(\alpha)\, \mathbb{I}}$$

Linear term

$\mathbb{E}_{q(\beta)}(\beta)\, \omega^T X^T y$ ~~$\Rightarrow x^T \Sigma^{-1}\mu \Rightarrow y^T \Sigma_\omega^{-1} X\omega \xrightarrow{\text{transpose}} \omega^T X^T \Sigma_\omega^{-1} y$~~

We want to re-write this such that $\omega \to x$ for our new gaussian

$$\Rightarrow \omega^T\left[\mathbb{E}_{q(\beta)}(\beta) X^T y\right] \simeq x^T \Sigma_\omega^{-1}\mu_\omega \Rightarrow \Sigma_\omega^{-1}\mu_\omega = \mathbb{E}_{q(\beta)}(\beta) X^T y$$

9

Solve for μ_ω

$$\Rightarrow \Sigma_\omega^{-1}\mu_\omega = \mathbb{E}_{q(\beta)}(\beta) X^T y$$

$$\boxed{\mu_\omega = \Sigma_\omega \mathbb{E}_{q(\beta)}(\beta) X^T y}$$

*Note: while technically we could have just continued with the "full" (bulky) form of $q^*(\omega)$, "converting" it into a gaussian is useful both analytically and computationally.

Next, find $q^*(\alpha)$:

Go back to log joint: $\ln p(y, \omega, \alpha, \beta) = \ln p(y \mid \omega, \beta) + \ln p(\omega \mid \alpha) + \ln p(\alpha) + \ln p(\beta)$

terms w/ α = $\underbrace{\ln p(\omega \mid \alpha)}_{\mathcal{N}(\omega \mid 0, \alpha^{-1} \mathbb{I})} + \underbrace{\ln p(\alpha)}_{\Gamma(\alpha \mid a_0, b_0)}$

From pg. 6

$$\Rightarrow \underbrace{\ln \left[\frac{1}{(2\pi)^{d/2} |\alpha^{-1}\mathbb{I}|^{1/2}} \right]}_{} - \frac{1}{2}(\omega - \mu)^T [\alpha^{-1}\mathbb{I}]^{-1} (\omega - \mu) \Rightarrow \frac{D}{2}\ln\alpha - \frac{1}{2}(\omega-\mu)^T[\alpha^{-1}\mathbb{I}](\omega-\mu)$$

$$-\ln\left[(2\pi)^{d/2}|\alpha^{-1}\mathbb{I}|^{1/2}\right]$$

$$= \underbrace{-\ln(2\pi)^{d/2}}_{\text{const}} - \ln|\alpha^{-1}\mathbb{I}|^{1/2}$$

$$-\ln|\alpha^{-1}\mathbb{I}|^{1/2} = -\frac{1}{2}\ln|\alpha^{-1}\mathbb{I}|$$

Using $|k\bar{A}| = k^D |A|$

$$\Rightarrow -\frac{1}{2}\ln(\alpha^{-1})^D \cancel{|\mathbb{I}|}^{1} - \frac{1}{2}\ln\alpha^{-D} = \frac{D}{2}\ln\alpha$$

From pg. 6

$$\Rightarrow \frac{D}{2}\ln\alpha - \frac{\alpha}{2}\omega^T\omega$$

for $\Gamma(\alpha \mid a_0, b_0) = \underbrace{\frac{b_0^{a_0}}{\Gamma(a_0)}}_{u} \alpha^{a_0 - 1} \exp(-b_0 \alpha)$

$$\ln p(\alpha) = \underbrace{\ln u}_{\text{constant wrt } \alpha} + \ln \alpha^{a_0 - 1} - b_0 \alpha = (a_0 - 1) \ln \alpha - b_0 \alpha$$

Back to $q^*(\alpha) = \underbrace{\frac{D}{2} \ln \alpha - \frac{\alpha}{2} \omega^T \omega}_{\text{from } p(\omega \mid \alpha)} + \underbrace{(a_0 - 1) \ln \alpha - b_0 \alpha}_{\text{from } p(\alpha)}$

Take $\mathbb{E}$ wrt. ω

$$\Rightarrow \frac{D}{2} \ln \alpha - \frac{\alpha}{2} \mathbb{E}_{q(\omega)}[\omega^T \omega] + (a_0 - 1) \ln \alpha - b_0 \alpha$$

$$\Rightarrow \left(\frac{D}{2} + a_0 - 1\right) \ln \alpha - \left[\frac{1}{2} \mathbb{E}_{q(\omega)}[\omega^T \omega] - b_0\right] \alpha \longrightarrow \Gamma(\alpha \mid a_N, b_N)$$

Note similarity in terms of α.

where $a_N - 1 = \frac{D}{2} + a_0 - 1$ and $\boxed{b_N = \frac{1}{2} \mathbb{E}_{q(\omega)}[\omega^T \omega] - b_0}$

$$\boxed{a_N = \frac{D}{2} + a_0}$$

Next find $q^*(\beta)$:

$$\ln p(y,\omega,\alpha,\beta) = \ln p(y|\omega,\beta) + \underbrace{\ln p(\omega|\alpha) + \ln p(\alpha)}_{\text{constants w.r.t. } \beta} + \ln p(\beta)$$

$$q^*(\beta) = \mathbb{E}_{q(\omega)} \ln \underbrace{p(y|\omega,\beta)} + \ln p(\beta)$$

$$\ln \left[\frac{1}{(2\pi)^{D/2} |\beta^{-1}\mathbb{I}|^{1/2}} \exp\left[-\frac{1}{2}(y - X\omega)^T (\beta^{-1}\mathbb{I})^{-1} (y - X\omega) \right] \right]$$

constant w.r.t. β

$$\Rightarrow \cancel{-\frac{D}{2}\ln 2\pi} - \frac{1}{2}\ln|\beta^{-1}\mathbb{I}| - \frac{1}{2}(y - X\omega)^T (\beta^{-1}\mathbb{I})^{-1}(y - X\omega)$$

$$\Rightarrow -\frac{1}{2}\ln (\beta^{-1})^D \cancel{|\mathbb{I}|}^{1} - \frac{\beta}{2}(y - X\omega)^T(y - X\omega) = \underline{\frac{D}{2}\ln\beta - \frac{\beta}{2}(y - X\omega)^T(y - X\omega)}$$

$[\beta^{-1}\mathbb{I}]^{-1}$: Use $(kA)^{-1} = \frac{1}{k}A^{-1}$

$$\Rightarrow (\beta^{-1}\mathbb{I})^{-1} = \frac{1}{\beta^{-1}}\mathbb{I}^{-1} = \beta\mathbb{I}$$

$$(y - X\omega)^T(y - X\omega)$$

$$\Rightarrow [y^T - (X\omega)^T](y - X\omega) = (y^T - \omega^T X^T)(y - X\omega)$$

$$\Rightarrow y^T y - y^T X\omega - \underbrace{\omega^T X^T y}_{\text{constant} \therefore \text{transpose freely}} + \omega^T X^T X\omega$$

$$\Rightarrow y^T y - 2y^T X\omega + \omega^T X^T X\omega$$

$$\ln p(\beta) = \ln \Gamma(\beta | c_0, d_0) = \ln \left[\underbrace{\frac{d_0^{c_0}}{\Gamma(c_0)}} \beta^{c_0 - 1} \exp(-d_0\beta) \right]$$

$$\Rightarrow \underbrace{\ln v}_{\text{constant w.r.t. } \beta} + (c_0 - 1)\ln\beta - d_0\beta$$

$$q^*(\beta) = \underbrace{\mathbb{E}_{q(\omega)} \left[\frac{D}{2}\ln\beta - \frac{\beta}{2}(y - X\omega)^T(y - X\omega) \right]}_{\text{from } \ln p(y|\omega,\beta)} + \underbrace{(c_0 - 1)\ln\beta - d_0\beta}_{\text{from } \ln p(\beta)}$$

$$q^*(\beta) = \mathbb{E}_{q(w)}\left[\frac{D}{2}\ln\beta - \frac{\beta}{2}(y-Xw)^T(y-Xw)\right] + (c_0-1)\cancel{\beta}\ln\beta - d_0\beta$$

$$= \frac{D}{2}\ln\beta - \frac{\beta}{2}\mathbb{E}_{q(w)}\left[(y-Xw)^T(y-Xw)\right] + (c_0-1)\ln\beta - d_0\beta$$

group by β:

$$\Rightarrow \left(\frac{D}{2}+c_0-1\right)\ln\beta - \cancel{\beta}\left[\frac{1}{2}\mathbb{E}_{q(w)}\left[(y-Xw)^T(y-Xw)\right] + d_0\right]\beta$$

match to Γ dist

$$\boxed{c_N = \frac{D}{2}+c_0} \qquad \boxed{d_N = \frac{1}{2}\mathbb{E}_{q(w)}\left[(y-Xw)^T(y-Xw)\right] + d_0}$$

$$d_N = \frac{1}{2}\mathbb{E}_{q(w)}\left[y^Ty - 2y^TXw + w^TX^TXw\right] + d_0 \Rightarrow \frac{1}{2}\left[y^Ty - 2\mathbb{E}_{q(w)}(y^TXw) + \mathbb{E}_{q(w)}(w^TX^TXw)\right] + d_0$$

for $\mathbb{E}_{q(w)}(y^TXw)$, since $q^*(w)\sim\mathcal{N}(\mu_w, \Sigma_w) \Rightarrow \underline{y^TX\mu_w}$ since $\mathbb{E}_{q(w)}(y^TXw) = y^TX\underbrace{\mathbb{E}_{q(w)}(w)}_{\mu_w}$

for $\mathbb{E}_{q(w)}(w^TX^TXw)$, we also have to account for variance.

We start with 2 properties of the Trace (Tr) operator:

1) $Tr(c) = c \rightarrow$ constant, 2) $Tr(ABC) = Tr(CAB)$

Since w^TX^TXw is a scalar, we can wrap it in a trace and use the cyclic property to group w

$$\Rightarrow w^TX^TXw = Tr(w^TX^TXw) = Tr\left[w^T(X^TXw)\right] = Tr\left[(X^TXw)w^T\right] = Tr(X^TXww^T)$$

Next, we use the fact that both Tr and $\mathbb{E}$ are linear operators

$$\therefore \mathbb{E}_{q(\omega)} \operatorname{Tr}(X^T X \omega\omega^T) = \operatorname{Tr}\left[\mathbb{E}_{q(\omega)}(X^T X \omega\omega^T)\right] = \operatorname{Tr}\left[X^T X\, \mathbb{E}_{q(\omega)}(\omega\omega^T)\right]$$

To find $\mathbb{E}_{q(\omega)}(\omega\omega^T)$, we use: $\Sigma_\omega = \mathbb{E}\left[(\omega-\mu_\omega)(\omega-\mu_\omega)^T\right] = \mathbb{E}\left[(\omega-\mu_\omega)(\omega^T-\mu_\omega^T)\right]$

$$\Rightarrow \mathbb{E}\left[\omega\omega^T - \omega\mu_\omega^T - \underbrace{\mu_\omega \omega^T}_{\text{scalar.}} + \mu_\omega\mu_\omega^T\right] \Rightarrow \mathbb{E}\left[\omega\omega^T - 2\omega\mu_\omega^T + \mu_\omega\mu_\omega^T\right] = \Sigma_\omega$$

$$\Rightarrow \mathbb{E}(\omega\omega^T) - 2\,\mathbb{E}(\omega)\mu_\omega^T + \mathbb{E}(\mu_\omega\mu_\omega^T) = \mathbb{E}(\omega\omega^T) - 2\mu_\omega\mu_\omega^T + \mu_\omega\mu_\omega^T = \mathbb{E}(\omega\omega^T) - \mu_\omega\mu_\omega^T = \Sigma_\omega$$

$$\underline{\mathbb{E}(\omega\omega^T) = \Sigma_\omega + \mu_\omega\mu_\omega^T} \quad \therefore$$

$$\operatorname{Tr}\left[X^T X\, \mathbb{E}_{q(\omega)}(\omega\omega^T)\right] = \operatorname{Tr}\left[X^T X(\Sigma_\omega + \mu_\omega\mu_\omega^T)\right] = \operatorname{Tr}\left[X^T X \Sigma_\omega + X^T X \mu_\omega\mu_\omega^T\right]$$

$$\Rightarrow \operatorname{Tr}(X^T X \Sigma_\omega) + \operatorname{Tr}(X^T X \mu_\omega\mu_\omega^T) \xrightarrow[\text{cyclic property}]{} \operatorname{Tr}(X^T X \Sigma_\omega) + \operatorname{Tr}(\underbrace{\mu_\omega^T X^T X \mu_\omega}_{\text{scalar}})$$

$$\therefore$$

$$\underline{\mathbb{E}_{q(\omega)}(X^T X \omega\omega^T) = \operatorname{Tr}(X^T X \Sigma_\omega) + \mu_\omega^T X^T X \mu_\omega}$$

$$d_N = \frac{1}{2}\left[y^T y - 2\,\mathbb{E}_{q(\omega)}(y^T X\omega) + \mathbb{E}_{q(\omega)}(\omega^T X^T X \omega)\right] + d_0$$

$$\boxed{d_N = \frac{1}{2}\left[y^T y - 2y^T X \mu_\omega + \operatorname{Tr}(X^T X \Sigma_\omega) + \mu_\omega^T X^T X \mu_\omega\right] + d_0}$$

Put Coordinate Ascent Variational Inference scheme together;

for $q^*(\omega)$: $\Sigma_\omega^{-1} = \mathbb{E}_{q(\beta)}(\beta) X^T X + \mathbb{E}_{q(\alpha)}(\alpha) \mathbb{1}$ (from pg. 8)

since $\alpha \sim \Gamma(a_0, b_0)$, $\beta \sim \Gamma(c_0, d_0)$ $\therefore$ ~~$\mathbb{E}(\alpha) = \frac{a_0}{b_0}$ and $\mathbb{E}(\beta) = \frac{c_0}{d_0}$~~

$\mathbb{E}(\alpha) = \frac{a}{b}$, $\mathbb{E}(\beta) = \frac{c}{d}$

$$\mu_\omega = \Sigma_\omega \mathbb{E}_{q(\beta)}(\beta) X^T y \quad \text{(from pg. 9)}$$

for $q^*(\alpha)$: $a_N = \frac{D}{2} + a_0$, $b_N = \frac{1}{2} \mathbb{E}_{q(\omega)}[\omega^T \omega] - b_0$ (from pg. 10)

$$= \frac{1}{2}[\Sigma_\omega + \mu_\omega^T \mu_\omega] - b_0 \quad \text{(from pg. 13)}$$

for $q^*(\beta)$: $c_N = \frac{D}{2} + c_0$, $d_N = \frac{1}{2}[y^T y - 2y^T X \mu_\omega + Tr(X^T X \Sigma_\omega) + \mu_\omega^T X^T X \mu_\omega] + d_0$ (from pg 13)

$$= \frac{1}{2}[\|y - X\omega\|^2 + Tr(X^T X \Sigma_\omega)] + d_0$$

Q.E.D.