Hyperparameter tuning using Evidence maximization

~~P(data|w)~~

finding evidence directly would be complex. It will involve solving of the integral which is not trivial ie

$$\int_w P_{prior}(w)\, P(data|w)\, dw$$

this likelihood ~~has standard form~~ contains term involving $\sigma(w^T x)$ because of which will be very hard to compute.

→ Evidence approximation

We can approximate the posterior distribution to get approximate posterior. This also gives us a way to calculate the evidence

Evidence ~~approx~~ $\approx f(w_{MAP})\, (2\pi)^{d/2}\, |\Sigma|^{1/2}$

where $f(w_{MAP})$ is posterior at w_{MAP}

this equation comes from solving approximate posterior which is distributed as ~~N(w|w_MAP, Σ^-1)~~ $N(w|w_{MAP}, \Sigma^{-1})$

where $\Sigma = (\Phi^T R \Phi + \alpha I)^{-1}$

⇒ Thus, evidence $\approx f(w_{MAP})\, (2\pi)^{d/2}\, |(\Phi^T R \Phi + \alpha I)^{-1}|^{1/2}$

$\log Ev = \log f(w_{MAP}) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma|$

Lets look at $f(w_{MAP})$ →

$f(w_{MAP}) \propto$ ~~p~~ prior $|_{w_{MAP}}$ * likelihood $|_{w_{MAP}}$

$\log(f(w_{MAP})) = C + \log(\text{prior})_{w_{MAP}} + \log \text{likelihood}|_{w_{MAP}}$

Now ~~log likelihood = ...~~

likelihood for logistic regression is ~~π~~ $\prod_i y_i^{t_i}(1-y_i)^{t_i}$

where $y_i = \sigma(w^T \phi(x_i))$

→ this likelihood function doesnt contain α.

Similarly likelihood for poisson regression → does not contain α.
and original regression

→ $\boxed{\frac{\partial \log \text{likelihood}}{\partial \alpha} = 0}$

using this function, lets maximise & reduce.

$$\log E = \log f(w_{MAP}) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log(|\Sigma|)$$

$$\frac{\partial \log E}{\partial \alpha} = \frac{\partial}{\partial \alpha}(\log f(w_{MAP})) + \frac{1}{2}\frac{\partial}{\partial \alpha}\log(|\Sigma|)$$

$$= \frac{\partial}{\partial \alpha}\left(C + \log(\text{prior})|_{w_{MAP}} + \log \text{likelihood}\right) + \frac{1}{2}\frac{\partial}{\partial \alpha}(\log|\Sigma|)$$

$$= \frac{\partial}{\partial \alpha}\left[C + \frac{d}{2}\log(1/2\pi) + \frac{d}{2}\log(\alpha) - \frac{1}{2}\alpha\|w_{MAP}\|^2 + \log \text{likelihood}\right) + \frac{1}{2}\frac{\partial}{\partial \alpha}\log(|\Sigma|)$$

$$= \frac{d}{2\alpha} - \frac{1}{2}\|w_{MAP}\|^2 + \frac{1}{2}\frac{\partial}{\partial \alpha}(\log|\Sigma|)$$

Now solve this

$\frac{\partial}{\partial \alpha} \log(|\Sigma|)$

Now $|\Sigma|$ = product of eigen values.

· product of eigen values of $(Q^T R Q + \alpha I)^{-1}$

If d_i are eigen values of $Q^T R Q$

$\Rightarrow$ $d_i + \alpha$ are eigen values of $(Q^T R Q + \alpha I)$

$\Rightarrow$ $\frac{1}{d_i + \alpha}$ are eigen values of $(Q^T R Q + \alpha I)^{-1}$

$\rightarrow$ $\frac{\partial}{\partial \alpha} \left[\log \left(\pi \frac{1}{d_i + \alpha} \right) \right]$

$= \frac{\partial}{\partial \alpha} \left(\sum_i \log \frac{1}{(d_i + \alpha)} \right)$

$= \frac{\partial}{\partial \alpha} \sum_i - \log(d_i + \alpha)$

$= - \frac{\partial}{\partial \alpha} \left[\sum_i \log(d_i + \alpha) \right]$

$= \sum_i \frac{-1}{d_i + \alpha}$

$\Rightarrow$ $\frac{\partial \log E}{\partial \alpha} = \frac{d}{2\alpha} - \frac{1}{2} \|W_{MAP}\|^2 - \frac{1}{2} \sum \frac{1}{d_i + \alpha} = 0$

$= \frac{d}{\alpha} - \|W_{MAP}\|^2 = \sum \frac{1}{d_i + \alpha}$

$\Rightarrow d - \sum \frac{\alpha}{d_i + \alpha} = \alpha \|W_{MAP}\|^2$

$\rightarrow \sum 1 - \frac{\alpha}{d_i + \alpha} = \alpha \|W_{MAP}\|^2 \Rightarrow \sum \frac{d_i}{d_i + \alpha} = \alpha \|W_{MAP}\|^2$

$$\Rightarrow \alpha = \frac{\frac{\sum \beta_i}{\partial \alpha_i}}{\frac{\sum \frac{d_i}{\alpha_i + \alpha}}{\|w_{MAP}\|^2}}$$

$$\alpha = \frac{\gamma}{\|w_{MAP}\|^2}$$

Now α can be optimised using this iterative process.

1) Initialise α
2) find $\|w_{MAP}\|^2$ using that α.
3) update α
4) Repeat until convergence.

Note that this α will be used to calculate gradient and Hessian.
→ Hence $\|w_{MAP}\|$ depends on α.

To calculate optimal s and θ_j in residual :→

$$\log E = \log(f(w_{MAP})) + d/2 \log(2\pi) + 1/2 \log(|\Sigma|)$$

$$\frac{\partial \log E}{\partial s} = \frac{\partial}{\partial s}\left[c + \frac{d}{2}\log(1/2\pi) + \frac{d}{2}\log(\alpha) - 1/2\|w_{MAP}\|^2 + \log \text{likelihood}\right]$$

$$+ \frac{\partial}{\partial s}\log(|\Sigma|)$$

$$= \frac{\partial}{\partial s}\log \text{likelihood} + 1/2 \frac{\partial}{\partial s}\log(|\Sigma|)$$

$$= \frac{\partial}{\partial s}\left[\sum_i \sum_j t_{ij}\left[\ln(y_{ij} - y_{i(j-1)})\right]\right] + 1/2 \frac{\partial}{\partial s}\log(|\Sigma|)$$

$$= \sum_i \sum_j (\theta_j - a)\left[\sigma(s(\theta_j - a))\right]\left[1 - \sigma(s(\theta_j - a))\right]$$
$$- (\theta_{j-1} - a)\left[\sigma(s(\theta_{j-1} - a))\right]\left[1 - \sigma(s(\theta_{j-1} - a))\right] + 1/2 \frac{\partial}{\partial s}\log(|\Sigma|)$$

~~This can be solved using gradient descent to find optimal~~

Now $\frac{\partial}{\partial s}(\log|\Sigma|)$

$\Sigma^{-1} = Q^T R Q + \alpha I$

$\rightarrow \frac{\partial}{\partial s} \log[|\Sigma|] = \frac{\partial}{\partial s} \log\left[\prod_i \frac{1}{d_i + \alpha}\right]$ where d_i are eigenvalues of $Q^T R Q$

$= -\frac{\partial}{\partial s} \sum_i \log(d_i + \alpha)$

$= -\sum_i \frac{1}{d_i + \alpha} \frac{\partial d_i}{\partial s}$

we can calculate s where $\frac{\partial \log EV}{\partial s} = 0$ using gradient descent.

and then use that s to calculate w_{MAP} and repeat this process.