# Expectation Maximization

Linear Regression
(Bayesian)

**Likelihood**

$$p(t \mid \omega, \underline{X}, \beta) = N(t \mid \Phi^T \omega, \beta^{-1} I)$$

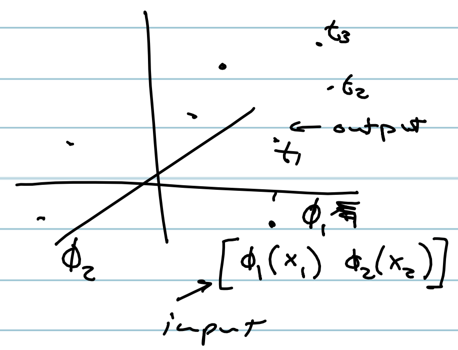
$$\omega^T = [\omega_1, \ldots, \omega_p] \qquad p \text{ is \# features}$$

$$\Phi = p \left\{ \begin{bmatrix} \phi_1(x_1) & \cdots & \phi_1(x_N) \\ \phi_2(x_1) & \cdots & \phi_2(x_N) \\ & & \\ \phi_p(x_1) & \cdots & \phi_p(x_N) \end{bmatrix} \right.$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{N}$$

**Prior**

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




Observed vars : $t, \underline{X}$ 's
Latent vars : $\omega$
Parameters : $\alpha, \beta$

Goal : estimate $\alpha, \beta$ in a M.L. using EM.

Analytic posterior for $\omega$
ie. $p(\omega \mid t, \underline{X}, \alpha, \beta)$

Using 2.3.3

$$p(x) = N(x \mid \mu, \Lambda^{-1})$$
$$p(y \mid x) = N(y \mid Ax + b, L^{-1})$$

$$\Rightarrow \quad p(y) = N(y \mid A\mu + b, L^{-1} + A \Lambda^{-1} A^T)$$
$$p(x \mid y) = N(x \mid \Sigma \{ A^T L (y - b) + \Lambda \mu \}, \Sigma)$$

where

$$\Sigma = (\Lambda + A^T L A)^{-1}$$

$$p(w \mid t, X, \alpha, \beta) = \mathcal{N}(w \mid S_n \{ \Phi \beta I (t) \}, S_n)$$

where

$$S_n = (\alpha I + \Phi \beta I \Phi^T)^{-1}$$

simplify

$$p(w \mid t, X, \alpha, \beta) = \mathcal{N}(w \mid \underbrace{\beta S_n \Phi t}_{p \times 1}, \underbrace{S_n}_{p \times p})$$

$$S_n^{-1} = \alpha I_p + \beta \Phi \Phi^T$$

## General EM Algorithm

Given joint dist $p(X, Z \mid \theta)$ over observed vars $X$ and latent vars $Z$, governed by params $\theta$.

Goal: maximize $p(X \mid \theta)$ w.r.t. $\theta$

[ remember $p(X \mid \theta) = \sum_Z p(X, Z \mid \theta)$

1) Choose init params $\theta^{old}$
2) E step Evaluate $p(Z \mid X, \theta^{old})$ ← posterior dist. of $w$
3) M step Evaluate $\theta^{new}$ given by

$$\theta^{new} = \arg\max_{\theta} Q(\theta, \theta^{old})$$

where

$$Q(\theta, \theta^{old}) = \sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta)$$

4) Check for convergence of either log like. or param values. If not converged

$$\theta^{old} \leftarrow \theta^{new}$$

goto 2).

$$\theta = \{\alpha, \beta\} \qquad Z = \omega \qquad X = t, X, \Phi$$

M step $\quad Q(\theta, \theta^{old}) = \underbrace{\mathbb{E}_{p(\omega|t,\Phi,\alpha,\beta)}\left[\ln p(t, \omega | \alpha, \beta)\right]}$

Maximize $Q$ w.r.t. $\alpha$ and $\beta$.

$$\mathbb{E}[\ln p(t, \omega|\alpha,\beta)] = \mathbb{E}[\ln (p(t|\omega,\cancel{\alpha},\beta)\, p(\omega|\alpha))]$$

$$= \mathbb{E}[\ln(\underbrace{p(t|\omega,\beta)}_{\text{Normal}})] + \mathbb{E}[\underbrace{\ln p(\omega|\alpha)}_{\text{Normal}}]$$

Remember

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

$$= \mathbb{E}[\ln N(t|\Phi^T\omega, \beta^{-1}I)] + \mathbb{E}[\ln N(\omega|0, \alpha^{-1}I)]$$

$\alpha$ and $\beta$ can be maxed independently

$$= \mathbb{E}\left[-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(|\beta^{-1}I|) - \frac{\beta}{2}(t-\Phi^T\omega)^T(t-\Phi^T\omega)\right]$$

$$+ \mathbb{E}\left[-\frac{P}{2}\ln(2\pi) - \frac{1}{2}\ln(|\alpha^{-1}I|) - \frac{\alpha}{2}(\omega^T\omega)\right]$$

[ Remember $|cI| = c^D$ when $I$ is a $D$-dim identity matrix

$$= \mathbb{E}\left[-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\beta^{-N} - \frac{\beta}{2}(t-\Phi^T\omega)^T(t-\Phi^T\omega)\right]$$

$$+ \mathbb{E}\left[-\frac{P}{2}\ln(2\pi) - \frac{1}{2}\ln\alpha^{-P} - \frac{\alpha}{2}(\omega^T\omega)\right]$$

$$= \frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\mathbb{E}\left[(t-\Phi^T\omega)^T(t-\Phi^T\omega)\right]$$

$$+ \frac{P}{2}\ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}\mathbb{E}[\omega^T\omega]$$

Want to maximize w.r.t. $\alpha$ & $\beta$

Standard approach, focus on $\alpha$

$$\frac{\partial}{\partial \alpha} \mathbb{E}\left[ \ln p(t, w \mid \alpha, \beta) \right] = 0 \qquad \text{solve}$$

$$\frac{P}{\cancel{2}} \cancel{2\pi} \frac{1}{\beta} \frac{1}{\cancel{2\pi}} - \frac{1}{\cancel{2}} \mathbb{E}[w^T w] = 0$$

$$\frac{P}{\beta} = \mathbb{E}[w^T w]$$

$$\Rightarrow \beta = \frac{P}{\mathbb{E}[w^T w]}$$

take as given, looked up

What is $\mathbb{E}[x^T x]$ where $x \sim \mathcal{N}(m, S)$

Consult Roweis' Gaussian cheat sheet

$$\int_x \underbrace{(x-\mu)^T \Sigma^{-1} (x-\mu)}_{\text{quadratic form}} \underbrace{\mathcal{N}(x \mid m, S)}_{\text{expectation}} \, dx$$

$$= (\mu - m)^T \Sigma^{-1} (\mu - m) + \mathrm{Tr}\left[\Sigma^{-1} S\right]$$

Simpler

$$\mathbb{E}[x^T x] = ? \,, \quad x \sim \mathcal{N}(m, S)$$
$$= m^T m + \mathrm{Tr}[S]$$

So $\mathbb{E}[w^T w] = ?$ $\quad w \sim \mathcal{N}(\beta S_n \Phi t, S_n)$

where $S_n^{-1} = \alpha I + \beta \Phi \Phi^T$

$$= (\beta S_n \Phi t)^T (\beta S_n \Phi t) + \mathrm{Tr}[S_n]$$

Call $m_n = \beta S_n \Phi t$, then at optimum $Q\{\{\alpha, \beta\}, \{\alpha, \beta\}^{old}\}$

$$\boxed{\alpha = \frac{P}{m_n^T m_n + \mathrm{Tr}[S_w]}} \leftarrow$$ analytic optimum of $\alpha$ in M step

- first M step procedure

Remember

from E step $\left\{ \begin{array}{l} m_n \text{ is mean of } p(w \mid t, \Phi, \alpha, \beta) \leftarrow \text{compute} \\ S_n \text{ is cov } \quad " \quad\quad " \quad\quad " \end{array} \right.$

gives dist $p(w \mid \cdots)$

$$\frac{\partial}{\partial\beta} \mathbb{E}[\ln p(t, w \mid \alpha, \beta)] =$$

$$\frac{\partial}{\partial\beta}\left[\frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\mathbb{E}\left[(t-\Phi^T w)^T(t-\Phi^T w)\right]\right] = 0$$

solve for $\beta$, opt. val. in M-step

$$\Rightarrow \frac{N}{2}\frac{2\pi}{\beta}\frac{1}{2\pi} - \frac{1}{2}\mathbb{E}\left[(t-\Phi^T w)^T(t-\Phi^T w)\right] = 0$$

$$\Rightarrow \frac{N}{2\beta} = \frac{1}{2}\mathbb{E}\left[(t-\Phi^T w)^T(t-\Phi^T w)\right]$$

$$\Rightarrow \beta = \frac{N}{\mathbb{E}\left[(t-\Phi^T w)^T(t-\Phi^T w)\right]}$$

using Roweis cheat sheet

$$\mathbb{E}\left[(t-\Phi^T w)^T(t-\Phi^T w)\right] \longleftarrow \text{incorrect before}$$

expectation over $w \sim N(m_n, S_n)$

$$= \cancel{(\Phi^T w)^T(\Phi^T w) + Tr[S_n]}$$

$$\boxed{\cancel{\Rightarrow \beta = \frac{N}{w^T\Phi\Phi^T w + Tr[S_n]}}} \quad \beta \text{ M-step update}$$

EM for linear regression (Bayesian)

E step:

compute

$$\left.\begin{aligned} m_n &= \beta S_n \Phi t \\ S_n^{-1} &= \alpha I + \beta \Phi\Phi^T \end{aligned}\right\} \text{Note, given } \alpha \text{ and } \beta$$

M step:

compute

$$\alpha = \frac{P}{m_n^T m_n + Tr[S_n]}$$

$$\cancel{\beta = \frac{N}{\underbrace{w^T}\Phi\Phi^T \underbrace{w} + Tr[S_n]}}$$

$w$ cannot appear

$$\mathbb{E}\left[ (t - \Phi^T w)^T (t - \Phi^T w) \right]$$

$$= \mathbb{E}\left[ t^T t - \underbrace{t^T \Phi^T w - (\Phi^T w)^T t}_{} + (\Phi^T w)^T (\Phi^T w) \right]$$

scalars, transpose then equil

$$\left( (\Phi^T w)^T t \right)^T = t^T \Phi w$$

$$= t^T t - 2 t^T \Phi^T \underbrace{\mathbb{E}[w]}_{\underbrace{\beta_{old} S_n \Phi t}_{m_n}} + \mathbb{E}\left[ w^T \Phi \Phi^T w \right]$$

in Roweis form

$$= \underbrace{t^T t - 2 t^T \Phi^T \beta_{old} S_n \Phi t \quad (-\beta_{old} S_n \Phi t)^T \Phi \Phi^T (-\beta_{old} S_n \Phi t)}_{\text{Quadratic form in } m_n} + \mathrm{Tr}\left[ \Phi \Phi^T \cdot S_n \right]$$

$$= (t - \Phi^T m_n)^T (t - \Phi^T m_n) + \mathrm{Tr}\left[ \Phi \Phi^T S_n \right]$$

$$\beta = \frac{N}{(t - \Phi^T m_n)^T (t - \Phi^T m_n) + \mathrm{Tr}\left[ \Phi \Phi^T S_n \right]}$$

$m_n$ and $S_n$ involve $\beta_{old}$ & pot. $\alpha_{old}$.

$\mathbb{B}$